

# $p$ -TORSION MONODROMY REPRESENTATIONS OF ELLIPTIC CURVES OVER GEOMETRIC FUNCTION FIELDS

BENJAMIN BAKKER AND JACOB TSIMERMAN

ABSTRACT. Given a complex algebraic curve  $C$  and a non-isotrivial family  $\mathcal{E}$  of elliptic curves over  $C$ , the  $p$ -torsion  $\mathcal{E}[p]$  yields a monodromy representation  $\rho_{\mathcal{E}}[p] : \pi_1(C) \rightarrow \mathrm{GL}_2(\mathbb{F}_p)$ . We prove that if  $\rho_{\mathcal{E}}[p] \cong \rho_{\mathcal{E}'}[p]$  then  $\mathcal{E}$  and  $\mathcal{E}'$  are isogenous, provided  $p$  is larger than a constant depending only on the gonality of  $C$ . This can be viewed as a function field analog of the Frey–Mazur conjecture, which states that an elliptic curve over  $\mathbb{Q}$  is determined up to isogeny by its  $p$ -torsion Galois representation for  $p \geq 17$ . The proof relies on hyperbolic geometry and is therefore only applicable in characteristic 0.

## 1. INTRODUCTION

The Frey–Mazur conjecture<sup>1</sup>, originating in [MG78], states that for a prime  $p \geq 17$ , an elliptic curve over  $\mathbb{Q}$  is classified up to isogeny by its  $p$ -torsion, viewed as a Galois representation (or equivalently, as a finite flat group scheme). A natural generalization of this conjecture asserts that over a fixed number field  $K$  there is a uniform  $C_K$  such that for primes  $p > C_K$ , elliptic curves over  $K$  are classified up to isogeny by their  $p$ -torsion representation. Moreover, one can hope that  $C_K$  can be made to depend only on the degree of  $K$ .

Geometrically, there is a surface  $Z(p)$  that parameterizes pairs of elliptic curves  $(E, E', \varphi)$  together with an isomorphism  $\varphi$  of their  $p$ -torsion, and this surface is endowed with natural Hecke divisors  $H_m$  parametrizing points for which  $\varphi$  is induced by a cyclic isogeny of degree  $m$ . The Frey–Mazur conjecture is equivalent to the statement that for  $p \geq 17$ , all rational points of  $Z(p)$  lie on one of these divisors<sup>2</sup>. By work of Hermann [Her91] the surface  $Z(p)$  is of general type for  $p > 11$ , and so the Bombieri–Lang conjecture implies that there are only finitely many rational points on the complement of all the rational and elliptic curves in  $Z(p)$ . Hence, it seems natural to first consider the Frey–Mazur conjecture over function fields of curves.

Our main result is to prove such a theorem for  $K$  the function field of a complex curve. Namely, we show that families of elliptic curves over curves are classified up to isogeny by the monodromy action on their  $p$ -torsion for sufficiently large  $p$ . In fact, we prove the stronger statement that the constant  $C_K$  depends only on the gonality of  $K$ . Recall that the gonality of an algebraic curve is the lowest degree map to  $\mathbb{P}^1$ , which is precisely the analogue of the degree of a number field in the function field setting. Precisely, we show:

---

*Date:* March 28, 2014.

<sup>1</sup>See Fisher [Fis11] for a survey of the Frey–Mazur conjecture.

<sup>2</sup>Note that by a theorem of Mazur, it is only necessary to consider  $m \leq 163$ .

**Theorem 1** (see Theorem 31). *Let  $k$  be an algebraically closed field of characteristic 0. For any  $B > 0$ , there exists  $C_B > 0$  such that for any smooth quasiprojective curve  $U$  of gonality  $n < B$  and prime  $p > C_B$ , non-isotrivial elliptic curves over  $U$  are classified up to isogeny by their  $p$ -torsion group scheme.*

*Equivalently, choosing a basepoint  $u \in U$ , non-isotrivial elliptic curves  $E$  over  $U$  are classified up to isogeny by the monodromy representation of the fundamental group  $\pi_1(U, u)$  on the 2-dimensional vector space  $E_u[p]$ .*

We can restate the above theorem in a way that seems more immediately analogous with the usual Frey–Mazur conjecture:

**Theorem 2.** *With  $k, B, C_B$  as above and for any smooth projective curve  $C$  of gonality  $n < B$ , non-isotrivial elliptic curves  $E$  over the field  $k(C)$  of rational functions on  $C$  are classified up to isogeny by their  $p$ -torsion Galois representations for  $p > C_B$ .*

Note that since the gonality of modular curves gets large [Abr96], for large enough  $p$  the Galois representations are all surjective onto  $\mathrm{SL}_2(\mathbb{F}_p)$ <sup>3</sup> and hence geometrically irreducible, so we don't have to worry about semi-simplifying the representations.

The statement of Theorem 1 is equivalent to the assertion that any map from a curve of gonality  $n < B$  to the modular surface  $Z(p)$  lies in a Hecke divisor  $H_m$  for  $p > C_B$ . However, it is easier to study curves in the product  $X(p) \times X(p)$ , where  $X(p)$  parameterizes elliptic curves together with an isomorphism  $E[p] \cong \mathbb{F}_p^2$ . The surface  $Z(p)$  is naturally the quotient of  $X(p) \times X(p)$  only remembering the composition  $E_1[p] \xrightarrow{\cong} \mathbb{F}_p^2 \xrightarrow{\cong} E_2[p]$ .

The main idea of the proof of Theorem 1 runs as follows: given a curve  $V$  of gonality  $n < B$  in  $Z(p)$ , we first get a genus 0 curve  $V'$  in  $\mathrm{Sym}^n Z(p)$ . We then lift it to a curve  $C$  in the variety  $(X(p) \times X(p))^n$  and estimate its genus using Riemann–Hurwitz in 2 different ways. We obtain a lower bound from the projections to the curves  $X(p)$  simply by ignoring ramification.

The upper bound requires a bound on the ramification of  $C \rightarrow V'$ , which can only be supported at the ramification points of the quotient map

$$(X(p) \times X(p))^n \rightarrow \mathrm{Sym}^n Z(p)$$

so we look to bound the number of times  $C$  can pass through this set. This constitutes the heart of the paper.

Our proof heavily relies on hyperbolic geometry and the fact that  $(X(p) \times X(p))^n$  is uniformized by the  $2n$ -th power  $\mathbb{H}^{2n}$  of the upper halfplane. The strategy is to bound the multiplicity of curves  $C$  along geodesic subvarieties of  $(X(p) \times X(p))^n$  in terms of their volume in small tubular neighborhoods of those subvarieties. A classical result of Federer states that a curve passing through the center of a ball  $B$  of radius  $r$  in  $\mathbb{C}^n$  must have volume in  $B$  at least equal to that of a coordinate axis. This result was generalized heavily by Hwang and To [HT02, HT12] to the case of arbitrary symmetric domains and higher-dimensional subvarieties. For our needs, the theorems of Hwang and To are not

<sup>3</sup>As we are working over complex curves, the Weil pairing is invariant under the monodromy action, so the representation lies in  $\mathrm{SL}_2$  rather than  $\mathrm{GL}_2$ .

quite sufficient, so we prove several analogues of these results, which may be interesting in their own right.

The bounds we obtain on the multiplicities of  $C$  along a subvariety are better for large radius neighborhoods, but in order to bound the multiplicity along many such subvarieties simultaneously it is necessary to understand how these neighborhoods overlap. We prove that special subvarieties tend to grow farther apart as  $p$  gets large, and that these subvarieties only “clump” together near higher-dimensional special subvarieties. The proofs of these repulsion results are arithmetic in nature and fundamentally use that fact that the monodromy group of  $X(p)$  over  $X(1)$  is an algebraic group.

We expect the methods here can be easily adapted to prove an analogue of Theorem 1 for abelian surfaces with quaternionic multiplication by replacing the modular curve with a compact Shimura curve following our preprint [BT13]<sup>4</sup>, but we do not pursue this here.

### 1.1. Outline of the paper.

In Section 2 we recall background on modular curves, including the modular interpretation of the compactifications  $X(p)$ , and introduce the basic structures on  $Z(p)$ . Our techniques require a uniformized metric on  $(X(p) \times X(p))^n$ , and in Section 3 we study the uniformized metric on  $X(p)$  in terms of the classical metric on  $Y(p)$ . Section 4 establishes the repulsion of special subvarieties of the product  $(X(p) \times X(p))^n$ , and in Section 5 we provide some machinery in the style of Hwang and To estimating the volume of curves in small neighborhoods of these subvarieties. Section 6 combines these results to provide estimates of the multiplicities of curves along special subvarieties, and in Section 7 we use this to estimate ramification and prove Theorem 1.

### 1.2. Acknowledgements.

The authors benefited from many useful conversations with Fedor Bogomolov, Johan de Jong, Michael McQuillan, Allison Miller, and Peter Sarnak. The first named author was supported by NSF fellowship DMS-1103982.

### 1.3. Notation and conventions.

Throughout the paper we use the following notation regarding asymptotic growth: for functions  $f, g$  we write  $f \gg g$  if there is a positive constant  $C > 0$  such that  $f - Cg$  is a positive function; likewise for  $\ll$ . If  $f_t, g_t$  are functions depending on  $t$ , we write  $f_t = O(g_t)$  if there is a positive constant  $C > 0$  such that  $C|g_t| - |f_t|$  is positive for  $t$  sufficiently large. If the same is true for any  $C > 0$  we write  $f_t = o(g_t)$ . We also write  $f_t = \omega(g_t)$  to mean  $g_t = o(f_t)$ . For us, the asymptotic parameter  $t$  will always be the prime  $p$ .

## 2. MODULAR CURVES

### 2.1. Basics on modular curves.

---

<sup>4</sup>In [BT13] the authors only prove the weaker result that the map from isogeny classes to  $p$ -torsion representations is 2 to 1. This can be rectified by using Proposition 26 as in Proposition 28.

For a prime number  $p > 3$  we let  $Y(p)$  denote the coarse moduli scheme representing pairs

$$(E, \varphi : \mathbb{F}_p^2 \xrightarrow{\cong} E[p])$$

of elliptic curves  $E$  together with a *projective* isomorphism  $\varphi$  from  $\mathbb{F}_p^2$  to the  $p$ -torsion of  $E$ —that is, an isomorphism  $\varphi : \mathbb{F}_p^2 \rightarrow E[p]$  defined up to scaling. We let  $X(p)$  denote the standard smooth compactification of  $Y(p)$ ; the added points  $X(p) - Y(p)$  are referred to as cusps.  $X(p)$  has 2 connected components, determined by the square class of the Weil pairing of  $\langle \varphi(e_1), \varphi(e_2) \rangle$ . We let  $X(p)_\epsilon$  denote the corresponding connected component, where  $\epsilon \in \mathbb{F}_p^\times / (\mathbb{F}_p^\times)^2$ . We shall consider these schemes exclusively over  $\mathbb{C}$ .

We recall that  $\mathrm{SL}_2(\mathbb{R})$  has a natural action on the upper half plane  $\mathbb{H}$  given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z = \frac{az + b}{cz + d}.$$

Letting  $\Gamma(p) := \{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \mid \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \pmod{p} \}$ , there is a natural isomorphism  $Y(p)_1 \simeq \Gamma(p) \backslash \mathbb{H}$  and for any  $\epsilon \in \mathbb{F}_p^\times$  there exist (several) isomorphisms  $Y(p)_1 \cong Y(p)_\epsilon$ . There is also a natural action of  $\mathrm{PGL}_2(\mathbb{F}_p)$  on  $Y(p)$  which permutes the two components, given by

$$g(E, \varphi) := (E, \varphi \circ \tilde{g}^{-1})$$

where  $\tilde{g}$  is any lift of  $g$  to  $\mathrm{GL}_2(\mathbb{F}_p)$ . Likewise there is an action of  $\mathrm{PGL}_2(\mathbb{F}_p)$  on  $X(p)$ , and the quotient is canonically the compactified modular curve  $X(1)$ ; call the quotient map  $\pi : X(p) \rightarrow X(1)$ . The ramification occurs at the cusps  $X(p) - Y(p)$  and at the pre-images under  $\pi$  of the points in  $i, \omega \in X(1)$  representing the elliptic curves  $E_i, E_\omega$  with CM by  $\mathbb{Z}[i]$  and  $\mathbb{Z}[\sqrt[3]{-1}]$  respectively. Each of these three sets forms a single orbit under  $\mathrm{PGL}_2(\mathbb{F}_p)$ , and the ramification order at a point in the orbit is  $p, 2$  and  $3$  respectively.

We remark that there is a natural anti-holomorphic involution on  $Y(1)$  given by negating the complex structure of the elliptic curve, and this induces an involution on  $Y(p)$  and  $X(p)$  as well. We denote this by  $z \rightarrow \bar{z}$ .

## 2.2. Modular interpretation of $n$ -gons.

$X(1)$  has several interpretations as the coarse space of moduli problems compactifying that of  $Y(1)$ ; usually this is done by considering the cusp point as the pointed nodal cubic or “1-gon,” but when considering elliptic curves with  $n$ -torsion it is more naturally thought of as the  $n$ -gon.

**Definition.** Let  $C$  be the nodal cubic and let  $C_n$  be the unique connected degree  $n$  étale cover of  $C$ . Geometrically  $C_n$  is a cyclic chain of  $\mathbb{P}^1$ s, obtained from  $\mathbb{Z}/n\mathbb{Z} \times \mathbb{P}^1$  by gluing  $(k, \infty)$  to  $(k+1, 0)$  at a node. An  $n$ -gon is  $C_n$  together with a group structure on the smooth locus  $C_n^{sm}$  such that the action of  $C_n^{sm}$  on  $C_n^{sm}$  extends to all of  $C_n$ . This latter requirement implies that  $C_n^{sm}$ —which is  $n$  copies of  $C^{sm}$ —is noncanonically  $\mathbb{Z}/n\mathbb{Z} \times \mathbb{G}_m$  as a group scheme, where  $\mathbb{Z}/n\mathbb{Z}$  acts by rotating the cycle of rational curves.

Note that the automorphism group of the  $n$ -gon is noncanonically  $\mathbb{Z}/n\mathbb{Z} \rtimes \mathbb{Z}/2\mathbb{Z}$ , the  $\mathbb{Z}/2\mathbb{Z}$  coming from inversion on the smooth locus. Indeed, choosing a

smooth point  $x \in C_n$  of order  $n$  not in the identity component  $(C_n^{sm})^0$  yields a group isomorphism

$$\mathbb{Z}/n\mathbb{Z} \times (C_n^{sm})^0 \xrightarrow{\cong} C_n^{sm} : (m, t) \mapsto tx^m$$

and each map  $(m, t) \mapsto (m, \zeta^m t)$  is a group automorphism of  $\mathbb{Z}/n\mathbb{Z} \times (C_n^{sm})^0$  for any  $n$ -th root of unity  $\zeta$ , yielding the  $\mathbb{Z}/n\mathbb{Z}$  of automorphisms.

$n$ -gons are useful because they can be endowed with full level  $n$  structure. Indeed, the  $n$ -torsion  $C_n^{sm}[n]$  is canonically Cartier self-dual, and thus there is a natural Weil pairing. We therefore define a level  $n$  structure of the  $n$ -gon  $C_n$  to be an isomorphism  $C_n^{sm}[n] \cong (\mathbb{Z}/n\mathbb{Z})^2$  up to scale. For the rest of the paper, for a fixed prime  $p$ , by a generalized elliptic curve we will mean either an elliptic curve or a  $p$ -gon, so that we may speak of generalized elliptic curves with full level  $p$  structure.

$X(p)$  can now be interpreted for  $p > 3$ ) as the fine moduli space of generalized elliptic curves with full level  $p$  structure. The stack  $\mathcal{X}(1)_p$  of generalized elliptic curves (in the above sense) without the level  $p$  structure has coarse space  $X(1)$  and compactifies the moduli problem  $\mathcal{Y}(1)$ ; it differs from  $\mathcal{X}(1) \cong \mathcal{X}(1)_1$  in that the cusp point in the  $\mathcal{X}(1)_p$  moduli problem has an order  $p$  stabilizer.

*Remark 3.* Throughout the above, we assume an algebraically closed base field, and thus blur the distinction between  $\mu_n$  and  $\mathbb{Z}/n\mathbb{Z}$ . Over a non-closed field, the  $n$ -torsion of  $C_n^{sm}$  is noncanonically a Cartier self-dual extension of  $\mathbb{Z}/n\mathbb{Z}$  by  $\mu_n$ , and the automorphism group of the  $n$ -gon is an extension of  $\mathbb{Z}/2\mathbb{Z}$  by  $\mu_n$ .

### 2.3. The diagonal quotient surface.

We now introduce our primary object of study.

**Definition.** Let  $Z(p)$  denote the coarse moduli scheme representing triples

$$(E_1, E_2, \psi : E_1[p] \xrightarrow{\cong} E_2[p])$$

consisting of a pair of generalized elliptic curves  $E_1, E_2$  together with a *projective* isomorphism between their  $p$ -torsion.

Note that there is a natural morphism  $X(p) \times X(p) \rightarrow Z(p)$  given by

$$(E_1, \varphi_1) \times (E_2, \varphi_2) \rightarrow (E_1, E_2, \varphi_2 \circ \varphi_1^{-1})$$

which identifies  $Z(p)$  with the quotient  $\mathrm{PGL}_2(\mathbb{F}_p) \backslash X(p) \times X(p)$  by the diagonal action of  $\mathrm{PGL}_2(\mathbb{F}_p)$ .

The surface  $Z(p)$  was introduced by Hermann [Her91] and studied by Kani–Schanz [KS98] and Carlton [Car01].

### 2.4. (anti-)Heegner CM points and singular bicusps.

Note that  $(x, y) \in X(p) \times X(p)$  is a ramification point of the quotient map  $X(p) \times X(p) \rightarrow Z(p)$  exactly if  $x$  and  $y$  share a common stabilizer in  $\mathrm{PGL}_2(\mathbb{F}_p)$ . Thus, either  $x, y$  are both cusps or are both in  $Y(p)$ .

Suppose that  $x, y \in Y(p)$  have a common stabilizer  $g \in \mathrm{PGL}_2(\mathbb{F}_p) - \mathbf{1}$ , so that they both map to either  $i$  or  $\omega$  in  $X(1)$ . Now, since  $g$  stabilizes  $x, y$  there must exist automorphisms  $h_x, h_y$  of  $E_x, E_y$  respectively and lifts  $g_x, g_y$  of  $g$  to  $\mathrm{GL}_2(\mathbb{F}_p)$  such that  $h_x \circ \varphi_x = \varphi_x \circ g_x^{-1}$  and  $h_y \circ \varphi_y = \varphi_y \circ g_y^{-1}$ . It is clear that neither of  $h_x, h_y$  are  $\pm 1$ . By possibly negating  $h_x, g_x$  we can ensure that  $h_x, h_y$  have the

same characteristic polynomial, either  $t^2 + 1$ ,  $t^2 - t + 1$  or  $t^2 + t + 1$ . It follows that we can take  $g_x = g_y$ .

**Definition.** Under the above setup, we say that  $(x, y) \in Y(p) \times Y(p)$  is a *Heegner CM point* if the eigenvalues of  $h_x$  acting on the tangent space  $T_0 E_x$  and  $h_y$  acting on  $T_0 E_y$  are the same, and an *anti-Heegner CM point* otherwise, in which case they are complex-conjugates. We denote these sets by  $\text{CM}^+, \text{CM}^- \subset X(p) \times X(p)$ , respectively.

It is easy to see that  $\text{PGL}_2(\mathbb{F}_p)$  preserves each the sets  $\text{CM}^+$  and  $\text{CM}^-$ . Note that  $(x, y)$  is a Heegner CM point if and only if  $(x, \bar{y})$  is an anti-Heegner CM point.

In the terminology of Kani and Schanz [KS98], a *Heegner CM point* of  $Z(p)$  is a point of the form  $(E, E, \psi|_{E[p]})$  for an elliptic curve  $E$  with  $\text{Aut}(E) \neq \pm \text{id}$  and  $\psi \in \text{End}(E)$  of degree coprime to  $p$ , whereas an *anti-Heegner CM point* is one of the form  $(E, \bar{E}, \tau \circ \psi|_{E[p]})$  for such  $E$  and  $\psi$ , where  $\tau : E \rightarrow \bar{E}$  is complex conjugation. Thus the (anti-)Heegner CM points of  $X(p) \times X(p)$  lie over (anti-)Heegner CM points  $(E, E, \varphi)$  of  $Z(p)$ .

The only other ramification points of the map  $X(p) \times X(p) \rightarrow Z(p)$  are those each of whose coordinates are cusps. We refer to these as *bicusp*s, and make the

**Definition.** A point  $(x, y) \in X(p) \times X(p)$  is called a *singular bicusps* if  $x, y$  are both cusps and they share a stabilizer in  $\text{PGL}_2(\mathbb{F}_p)$ . We denote the set of them by  $\text{SBC} \subset X(p) \times X(p)$ .

## 2.5. Hecke operators.

Recall that if we have an integer  $n$  relatively prime to  $p$ , we can define a Hecke correspondence  $T_n \subset X(p) \times X(p)$  between  $X(p)$  and itself as the closure of:

$$\{((E_1, \varphi_1), (E_2, \varphi_2)) \mid \exists \text{ cyclic isogeny } \psi : E_1 \rightarrow E_2, \deg \psi = n, \psi \circ \varphi_1 = \varphi_2\}.$$

We describe also the correspondence on  $p$ -gons with full level-structure. First, consider the map  $\varphi_n : C_{pn} \rightarrow C_p$  which on the smooth part is just

$$\varphi_n(x, a) = (x^n, a), (x, a) \in \mathbb{G}_m \times \mathbb{Z}/n\mathbb{Z}.$$

This induces an isomorphism on  $p$ -torsion, which we denote by  $\varphi_n[p]$ . Next, consider the set  $G_n$  of all subgroups  $G \subset C_{pn}^{sm}$  which are cyclic of order  $n$  and whose intersection with the identity component is a single point. For each such  $G \in G_n$  we get a map  $\psi_G : C_{pn}^{sm} \rightarrow C_p^{sm}$  by quotienting out by  $G$ , which completes to a map from  $C_{pn}$  to  $C_p$ . Finally, consider the map  $\xi_d : C_p \rightarrow C_p$  which is  $x \rightarrow x^d$  on the smooth part of the identity component, so that

$$\xi_d(x, a) = (x^d, a), (x, a) \in \mathbb{G}_m \times \mathbb{Z}/n\mathbb{Z}.$$

Now, to a point  $(C_p, f : \mathbb{F}_p^2 \cong C_p^{sm}[p])$  the  $n$ th Hecke operator associates the set

$$\bigcup_{m|n} \bigcup_{G \in G_m} (C_p, \xi_{\frac{n}{m}} \circ \psi_G \circ \varphi_m[p]^{-1} \circ f).$$

It will be important for us that  $T_n$  is (diagonally)  $G(p)$ -invariant. The two projection maps  $\alpha, \beta : T_n \rightarrow X(p)$  both have degree  $\deg(T_n) = \sigma_1(n) = \sum_{d|n} d$ , and we denote by  $\mu, \nu : X(p) \times T_n \rightarrow X(p) \times X(p)$  the maps  $\mu = \text{id} \times \alpha$  and

$\nu = \text{id} \times \beta$ .  $\mu, \nu$  yield a correspondence from  $X(p) \times X(p)$  to itself, and we denote by  $T_m^* = \mu_* \nu^*$  the pullback of divisors along this correspondence.

### 3. HYPERBOLIC PROPERTIES OF $X(p)$

The modular curve  $Y(p)$  carries a natural uniformized metric  $h_{Y(p)}$ .  $X(p)$  is a hyperbolic curve (for  $p > 3$ ) in and of itself, and thus also carries a uniformized metric  $h_{X(p)}$ . Though the classical metric  $h_{Y(p)}$  is well-understood, it's singularities at the cusps make it unamenable to the techniques of Hwang and To. The purpose of this section is to study the properties  $h_{X(p)}$  by comparing it with  $h_{Y(p)}$ .

The orbifold coarse space of the stack  $\mathcal{X}(1)_p$  of generalized elliptic curves is topologically a sphere with a point  $c_2$  of order 2, and point  $c_3$  of order 3, and a point  $c_p$  of order  $p$ . Let  $X(1)_p$  be the (orbifold) Riemann surface associated to the tiling of the upper halfplane by a  $(2, 3, p)$  triangle in the same way that  $Y(1)$  is associated to the tiling by the  $(2, 3, \infty)$  triangle. Let

$$\Gamma(2, 3, p) = \langle \sigma_2, \sigma_3, \sigma_p \mid \sigma_2^2 = \sigma_3^3 = \sigma_p^p = \sigma_2 \sigma_3 \sigma_p = 1 \rangle = \pi_1(X(1)_p)$$

be the  $(2, 3, p)$  triangle group. The unique biholomorphism from the usual fundamental domain of  $Y(1)$  to a  $(2, 3, p)$  triangle yields by Schwarz reflection the holomorphic embedding  $i_p : Y(1) \rightarrow X(1)_p$ , by which we identify  $X(1)_p$  with the orbifold coarse space of  $\mathcal{X}(1)_p$ . We let

$$\gamma_p := i_{p*} : \pi_1(Y(1)) \rightarrow \pi_1(X(1)_p)$$

After the usual identification  $\pi_1(Y(1)) = \text{PSL}_2 \mathbb{Z} \cong \Gamma(2, 3, \infty)$ ,  $\gamma_p$  is the obvious map  $\Gamma(2, 3, \infty) \rightarrow \Gamma(2, 3, p)$ . The forgetful map  $X(p) \rightarrow X(1)_p$  (of coarse moduli schemes) is identified with the étale cover associated to the image  $\Xi(p)$  of  $\Gamma(p)$  under  $\gamma_p$ , and the metric  $h_{X(p)}$  on  $X(p)$  is the pullback of the uniformized metric on  $X(1)_p$  coming from the above tiling. We define

$$G(p) := \Gamma(2, 3, p) / \Xi(p) \cong \text{PGL}_2(\mathbb{F}_p)$$

to be the Galois group of  $X(p)$  over  $X(1)_p$ .

#### 3.1. Injectivity radii.

For a Riemann surface  $X$  we will denote by  $h_X$  its uniformized metric of constant sectional curvature  $-1$ , and  $d_X$  the associated distance function. If  $\pi : \mathbb{H} \rightarrow X$  is the uniformizing map, then

$$\pi^* h_X = h_{\mathbb{H}} = \frac{dx^2 + dy^2}{y^2}$$

Recall that for a compact Riemann surface  $X$  endowed with  $h_X$ , the injectivity radius  $\rho_X(x)$  at a point  $x \in X$  is the largest radius for which the exponential map at  $x$  is a diffeomorphism. It is equal to half the length of the smallest closed geodesic through  $x$ . The injectivity radius  $\rho_X$  is the infimum of  $\rho_X(x)$  over all  $x \in X$ , or equivalently half the length of the shortest closed geodesic in  $X$ . If we now allow  $X$  to have cusps, we define  $\rho_X$  to be the infimum of the lengths of closed geodesics with respect to  $h_X$  that are homotopically nontrivial in the smooth compactification  $X'$ .



It was first observed by Buser-Sarnak in [BS94] that the injectivity radius of  $Y(p)$  with respect to the metric  $h_{Y(p)}$  grows. For the convenience of the reader, we recall the proof:

**Lemma 4.**  $\rho_{Y(p)} = 2 \log p + O(1)$ .

*Proof.* The kernel of  $\gamma_p$  is the group generated by the unipotents in  $\Gamma(p)$ . Thus, a homotopically nontrivial closed geodesic through  $x \in Y(p)$  lifts to the unique geodesic arc between two lifts  $z, \gamma z \in \mathbb{H}$  for some semisimple  $\gamma \in \Gamma(p)$ , so that  $d(z, \gamma z) = d_{\mathbb{H}}(Az, aAz)$ , where  $A \in \mathrm{SL}_2 \mathbb{R}$  is the diagonalizing matrix and  $\sqrt{a} + \frac{1}{\sqrt{a}} = |\mathrm{tr} \gamma|$ . In particular, using the formula for distance in the upper half-plane, this means

$$\begin{aligned} \min_z d_{\mathbb{H}}(z, \gamma z) &= \min_z d_{\mathbb{H}}(z, az) \\ &= \min_z \operatorname{arcosh} \left( 1 + \frac{(a-1)^2 |z|^2}{2a(\Im z)^2} \right) \\ &\geq 2 \log |\mathrm{tr} \gamma| \end{aligned}$$

The only semisimple element  $\gamma$  with trace 2 is the identity, and thus the minimal value of  $|\mathrm{tr} \gamma|$  for  $\mathbf{1} \neq \gamma \in \Gamma(p)$  is  $p^2 - 2$  as  $\mathrm{tr}(\gamma) \equiv 2 \pmod{p^2}$ . Moreover, this bound can be achieved by taking  $\gamma = \begin{pmatrix} 1-p^2 & p \\ 1 & -p \end{pmatrix}$ . The result then follows.  $\square$

### 3.2. Comparison of $h_{Y(p)}$ and $h_{X(p)}$ .

This subsection will show that the metrics  $h_{X(p)}$  and  $h_{Y(p)}$  can be compared away from the cusps, and that the metric  $h_{X(p)}$  on  $X(p)$  also has growing injectivity radius.

Let  $T_{2,3,p} \subset \mathbb{H}$  denote the  $(2, 3, p)$  triangle with vertices at  $i, iy_p, e^{i\theta_p}$ , with a right angle at  $i$  and  $0 < \theta_p < \pi/2$  and an angle of  $\frac{\pi}{3}$  at  $e^{i\theta_p}$ . We define  $\Delta_{2,3,p}$  to be the fundamental domain (see Figure 1) of the corresponding action of  $\Gamma(2, 3, p)$  on  $\mathbb{H}$  given by the union of  $T_{2,3,p}$  with its reflection through the imaginary axis. Likewise,  $\Delta_{2,3,\infty}$  is the usual fundamental domain for  $Y(1)$ . By the second hyperbolic law of cosines, we compute  $\cos(\pi/3) = \sin(\pi/p) \cosh(\log y_p)$  from which we conclude that  $y_p = p + O(\frac{1}{p})$ . Similarly, we conclude that  $\theta_p = \pi/6 + O(\frac{1}{p})$ .

Henceforth we will implicitly think of  $\Gamma(2, 3, p)$  as embedded in  $\mathrm{PSL}_2 \mathbb{R}$  via the tiling by  $T_{2,3,p}$  and its reflection.  $X(p)$  is then tiled by a set  $\Sigma$  of the images of these tiles, so that  $G(p)$  acts on  $\Sigma$  with 2 orbits. It is easy to see (from Figure 1, for instance) that

**Lemma 5.** *For any cusp  $c \in X(p)$ , the disk of radius  $\log p - 1$  centered at  $c$  intersects only those triangles in  $\Sigma$  with  $c$  as a vertex. As a consequence, if  $c, c'$  are two cusps, then  $d_{X(p)}(c, c') > 2 \log p - 2$ .*

Let  $f_p : \Delta_{2,3,\infty} \rightarrow \Delta_{2,3,p}$  be the unique map sending  $i$  to  $i$ ,  $\omega = e^{2\pi i/3}$  to  $e^{i\theta_p}$ , and  $\infty$  to  $iy_p$  that is a biholomorphism on the interiors of the two fundamental domains and continuous on the boundary. By Schwarz reflection, these maps glue together to a holomorphic map  $F_p : \mathbb{D} \rightarrow \mathbb{D}$  lifting the inclusion  $i_p : Y(1) \rightarrow X(1)_p$  to their universal covers; for convenience we've here used the Poincaré disk model, where

$$h_{\mathbb{D}} = \frac{4|dz|^2}{(1 - |z|^2)^2}$$



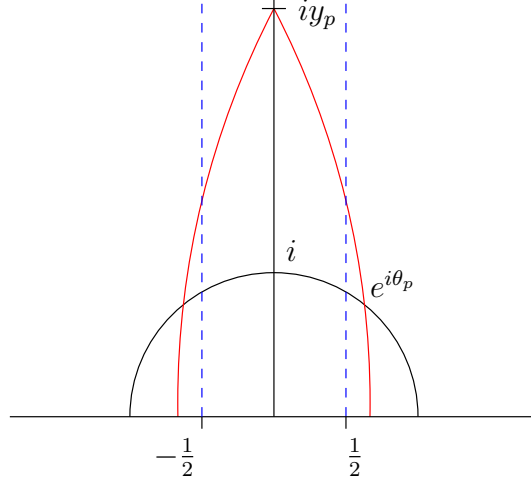


FIGURE 1. Fundamental domains of  $Y(1)$  and  $X(1)_p$ .  $\Delta_{2,3,\infty}$  is bordered by the arc of the unit circle together with the two dotted blue lines, while  $\Delta_{2,3,p}$  is bordered by the arc of the unit circle together with the two solid red lines.

so that

$$\tanh^2(d_{\mathbb{D}}(z, w)/2) = \left| \frac{w - z}{1 - \bar{z}w} \right|^2.$$

Note that  $F_p$  is in fact étale, though it will not be surjective onto  $\mathbb{D}$  (it will miss the preimages of the cusp  $c_p \in X(1)_p$ ). By the Schwarz–Pick theorem,  $F_p$  is distance decreasing with respect to the hyperbolic metric  $h_{\mathbb{D}}$ , so

$$F_p^* h_{\mathbb{D}} \leq h_{\mathbb{D}}$$

and therefore  $h_{X(1)_p}|_{Y(1)} \leq h_{Y(1)}$ . In fact, the two metrics are close far from the cusp:

**Proposition 6.** *For  $x \in X(1)_p$ , let  $d_{\text{cusp}}(x) = d_{X(1)_p}(x, c_p)$ . Then*

$$\tanh^2(d_{\text{cusp}}/2) h_{Y(1)} \leq h_{X(1)_p}|_{Y(1)} \leq h_{Y(1)}$$

*Proof.* Given  $x \in Y(1)$ , by pre- and post-composing  $F_p$  with hyperbolic isometries we may assume that  $x$  lifts to  $0 \in \mathbb{D}$  under both uniformizations, and that  $F_p(0) = 0$ . Let  $d = d_{\text{cusp}}(x)$ , and note that  $F_p^{-1}$  exists on the hyperbolic ball  $B(0, d)$ , which is the euclidean ball of radius  $\tanh(d/2)$ . By Schwarz’s lemma,  $(F_p^{-1})'(0) \leq 1/\tanh(d/2)$ , and so at  $z = 0$

$$F_p^* h_{\mathbb{D}} = 4|F_p'(0)|^2 |dz|^2 \geq 4 \tanh^2(d/2) |dz|^2 = \tanh^2(d/2) h_{\mathbb{D}}$$

yielding the claim. □

Lifting to  $X(p)$ , we have the following immediate corollary:

**Corollary 7.** *For  $x \in X(p)$ , if  $d_{\text{cusp}}(x)$  is the minimum distance to a cusp, then*

$$\tanh^2(d_{\text{cusp}}/2) h_{Y(p)} \leq h_{X(p)}|_{Y(p)} \leq h_{Y(p)}$$

We can also conclude that the injectivity radius of  $X(p)$  is close to that of  $Y(p)$ :

**Corollary 8.** *The injectivity radius  $\rho_{X(p)} = 2 \log p + O(1)$ .*

*Proof.* The upper bound follows by Lemma 4 combined with the fact that  $h_{X(p)} \leq h_{Y(p)}$ . For the lower bound, suppose  $\gamma$  is a minimal length geodesic loop in  $X(p)$ , of length  $\ell_{X(p)}(\gamma)$ . Note that  $\gamma$  can be within  $2 \log p/3$  of at most one cusp, as  $\ell_{X(p)}(\gamma) \leq 2 \log p + O(1)$ . Thus, as  $\gamma$  is a geodesic, the length of  $\gamma$  in the  $h_{X(p)}$  metric within a distance  $d < \log p/2$  of any cusp is at most  $2d$ . Consider a new loop  $\gamma'$  which is equal to  $\gamma$ , except if  $\gamma$  is ever within distance 1 of a cusp, that stretch is moved to the boundary of a ball of radius 1 around that cusp. Thus, by Lemma 7

$$\begin{aligned} \rho_{Y(p)} &\leq \ell_{Y(p)}(\gamma') \\ &\leq \ell_{X(p)}(\gamma') + (\ell_{Y(p)}(\gamma') - \ell_{X(p)}(\gamma')) \\ &\leq O(1) + \ell_{X(p)}(\gamma) + \int_{x=1}^{\infty} 2(1 - \tanh(x/2)) dx \\ &= \rho_{X(p)} + O(1) \end{aligned}$$

Thus the claim follows by Lemma 4.  $\square$

The function  $d_{cusp}(x)$  from Proposition 7 is closely related to the “imaginary height” notion of distance to the cusp in  $Y(1)$ . Indeed, switching back to thinking of  $Y(1)$  as uniformized by the upper halfplane  $\mathbb{H}$ , for  $x \in Y(1)$ , after lifting  $x$  to the unique point  $z \in \Delta_{2,3,\infty}$  we define

$$d_{im}(x) := \Im z$$

Note that the map  $F_p : \mathbb{H} \rightarrow \mathbb{D}$  factors through the punctured disk  $e^{2\pi iz/p} : \mathbb{H} \rightarrow \mathbb{D}^*$ . By Schwarz’s lemma applied to the resulting map  $G_p : \mathbb{D} \rightarrow \mathbb{D}$  we have

$$|G_p(z)| \leq |z|$$

Moreover, as in the proof of Lemma 6,  $G_p^{-1}$  is defined on a euclidean disk of radius  $r_p = 1 + O(1/p)$ , and so by Schwartz’s lemma again

$$|G_p(z)| = |z|(1 + O(1/p))$$

which means

$$e^{-2\pi \Im z/p} = \tanh(d_{cusp}(x)/2)(1 + O(1/p)).$$

We’ve thus proven the

**Lemma 9.** *With  $d_{im}$  and  $d_{cusp}$  defined as above, uniformly on  $Y(1)$  we have*

$$-\frac{2\pi d_{im}}{p} = \log \tanh(d_{cusp}/2) + O(1/p)$$

For a point  $x \in X(p)$ , denote by  $B_{X(p)}(x, R)$  the ball centered at  $x$  of radius  $R$  with respect to  $h_{X(p)}$ . We shall use the following simple corollary:

**Corollary 10.** *For any  $R < 2 \log p$ , any integer  $m > 0$ , and any cusp  $c \in X(p)$ , if  $T_m^* c = \cup_i c_i$  then*

$$T_m^* B_{X(p)}(c, R) \subset \bigcup_i B_{X(p)}(c_i, R + \log m + O(1)).$$

*Proof.* Note that  $T_m$  on the upper half plane uniformizing  $Y(p)$  satisfies  $\Im(T_m^* z) \geq \Im(z)/m$ . The claim thus follows immediately from Lemma 9 combined with the fact that for  $d > 1$ ,  $\log \tanh(d/2) = -2e^{-d+O(1)}$ .  $\square$

### 3.3. Heights.

Fix now a uniformizing map  $\mathbb{H} \rightarrow X(p)_1$ , and denote by  $H = \langle \sigma_p \rangle \subset \Gamma(2, 3, p)$  the stabilizer of  $iy_p$ , where  $\sigma_p$  is rotation by  $\frac{2\pi}{p}$  around  $iy_p$ . For convenience, we define the usual notion of height on  $\mathrm{SL}_2(\mathbb{Z})$ :

**Definition.** For  $M \in \mathrm{SL}_2(\mathbb{Z})$  denote by  $h(M)$  the maximum absolute value of its entries.

**Lemma 11.** Fix  $\iota \in X(p)_1$  to be the image of  $i \in \mathbb{H}$ , and suppose  $d_{g'}(\iota, \gamma \cdot \iota) = R$  where  $\gamma \in G(p)$ . Then there exists a matrix  $M \in \mathrm{SL}_2(\mathbb{Z})$  with  $h(M) = O(e^{2R})$  such that  $\gamma$  is the reduction of  $\gamma_p(M)$ .

*Proof.* By Corollary 7 it suffices to show in  $\mathbb{H}$  that if  $d_{\mathbb{H}}(i, Mi) = R$  for  $M \in \mathrm{SL}_2(\mathbb{Z})$  then the elements of  $M$  are  $O(e^{2R})$ .

Let

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

so that

$$Mi = \frac{ai + b}{ci + d} = \frac{(bd + ac) + i}{c^2 + d^2} \quad (1)$$

It follows by looking at the imaginary part of  $Mi$  that

$$c^2 + d^2 \leq e^R$$

and thus each of  $c, d$  are at most  $e^{R/2}$ . Using the distance formula on the upper half plane we have

$$2\Im(Mi)(\cosh(R) - 1) = (\Re Mi)^2 + (-1 + \Im Mi)^2$$

and thus

$$(\Re Mi)^2 \leq 2\Im(Mi)e^R$$

which from equation (1) gives

$$\frac{bd + ac}{c^2 + d^2} \leq \sqrt{2\Im(Mi)e^R}$$

and

$$bd + ac \leq 2e^{\frac{3R}{2}}$$

Now, using  $ad - bc = 1$  we get

$$a(c^2 + d^2) \leq 2ce^{\frac{3R}{2}} + d$$

so that

$$a \leq 2e^{2R}$$

and similarly for  $b$ .  $\square$

**Lemma 12.** Fix  $c_0 \in X(p)_1$  to be the image of  $iy_p \in \mathbb{H}$ .

- (a) For any cusp  $c \in X(p)_1$  and any lift  $z \in \mathbb{H}$  of  $c$  with  $d(iy_p, z) \leq (2 + \delta) \log p$ , there exists  $M \in \mathrm{SL}_2(\mathbb{Z})$  with  $h(M) = O(p^{2\delta})$  such that  $z = \sigma \gamma_p(M) \cdot iy_p$  for some  $\sigma \in H$ .
- (b) For all  $\gamma \in G(p)$ , if  $d(c_0, \gamma \cdot c_0) \leq (2 + \delta) \log p$  then there exists  $M \in \mathrm{SL}_2 \mathbb{Z}$  with  $h(M) = O(p^{2\delta})$  such that  $\gamma \in H \cdot \gamma_p(M) \cdot H$ .

*Proof.* By Lemma 5 we have  $d(iy_p, z) > 2 \ln p - 2$ , and so

$$B(iy_p, (1 + \delta) \log p) \cap B(z, (1 + \delta) \log p)$$

is within a distance of  $\delta \log p$  of the boundary of  $B(iy_p, \log p)$  as well as that of  $B(z, \log p)$ . Thus, it is within  $\delta \log p + O(1)$  of a point  $u \in \mathbb{H}$  which projects to  $c_2$  in  $X(1)_p$  and is a vertex of one of the  $(2, 3, p)$  tiles having  $iy_p$  as a vertex, so that  $u = \sigma \cdot i$  for some  $\sigma \in \langle \sigma_p \rangle$ . It thus follows by Lemma 11 that there is a matrix  $M \in \mathrm{SL}_2(\mathbb{Z})$  with  $h(M) = O(p^{2\delta})$  such that  $\gamma_p(M) \cdot i = \sigma^{-1} \cdot v$ , and by possibly pre-composing  $\gamma_p(M)$  with  $\sigma_2$  we get that  $\sigma \gamma_p(M) \cdot iy_p = z$ , thus proving (a).

Part (b) easily follows from part (a) after choosing a lift  $z \in H$  of  $\gamma \cdot c_0$  with  $d(c_0, \gamma \cdot c_0) = d(iy_p, z)$ . □

#### 4. REPULSION RESULTS

The volume estimates from Section 5 will allow us to bound the multiplicity of curves  $C$  in  $X(p) \times X(p)$  along special points in terms of their volume near those points. As  $p$  grows, the points tend to spread out further, and we can find neighborhoods of large radius around them that tend to be disjoint. The total volume, and therefore the total multiplicity, can therefore be bounded by the total volume of the curve. This argument fails when such neighborhoods overlap many times, but luckily such overlaps only occur close to higher dimensional special subvarieties, which themselves repel one another.

Throughout this section, we solely consider the metric  $h_{X(p)}$  from Section 3 on  $X(p)$ , and suppress its mention from the notation. Thus, the distance  $d_{X(p)}(x, x')$  between  $x, x' \in X(p)$  with respect to  $h_{X(p)}$  will be denoted  $d(x, x')$ , and the ball around  $x$  by  $B(x, R)$ . We use the same notation for distance and balls in  $\mathbb{H}$  as in  $X(p)$ , and rely on context to distinguish between the two.

##### 4.1. Repulsion of cusps.

**Proposition 13.** *For all sufficiently small  $\delta > 0$  and all sufficiently large  $p$ :*

- (a) For any distinct cusps  $c_0, c \in X(p)$ , and any pre-image  $z_0 \in \mathbb{H}$  of  $c_0$ , there is at most one pre-image  $z \in \mathbb{H}$  of  $c$  such that  $d(z_0, z) \leq (2 + \delta) \log p$ .
- (b) For any distinct cusps  $c_0, c \in X(p)$ ,

$$B(c_0, (1 + \delta) \log p) \cap B(c, (1 + \delta) \log p)$$

*is contained in a ball of radius  $\delta \log p + O(1)$ .*

- (c) For any  $x \in X(p)$  there are at most  $O(p^{12\delta})$  cusps  $c \in X(p)$  within a distance  $(1 + \delta) \log p$  of  $x$ .

*Proof.* To prove part (a), first note that because the automorphisms of  $X(p)$  act transitively on the cusps, it suffices to take  $z_0 = iy_p$ .

Let  $z, z' \in \mathbb{H}$  be pre-images of cusps  $c, c'$  with  $d(iy_p, z) \leq 2 \log p + 2R$ , and likewise for  $z'$ . We'll first show that  $c \neq c'$ . By Lemma 12 part (a), there are

matrices  $M, M' \in \mathrm{SL}_2(\mathbb{Z})$  with  $h(M), h(M') = O(p^{2\delta})$  and elements  $\sigma, \sigma' \in \langle \sigma_p \rangle$  such that  $\sigma \gamma_p(M) \cdot iy_p = z$  and  $\sigma' \gamma_p(M') \cdot iy_p = z'$ .

Since  $z, z'$  are translates under  $\Xi(p)$ , then  $\gamma_p(MM'^{-1}) \in \Xi(p)\langle \sigma_p \rangle$ . Since the kernel of  $\gamma_p$  is contained in  $\Gamma(p)$  we have that  $MM'^{-1} \in \Gamma(p)U$ , where  $U$  is the upper triangular group. In other words, the reduction of  $MM'^{-1}$  modulo  $p$  is upper triangular. However, as the entries of  $MM'^{-1}$  are  $O(p^{4\delta})$ , it follows for  $\delta < \frac{1}{4}$  that for large enough  $p$  we must have  $MM'^{-1} \in U$ , which implies that  $\gamma_p(MM'^{-1}) \in \langle \sigma_p \rangle$ , and

$$\sigma \gamma_p(MM'^{-1}) \sigma'^{-1} \in \Xi(p) \cap \langle \sigma_p \rangle = 1.$$

Thus we must have  $z = z'$ . This proves (a).

To prove part (b), after fixing a lift  $z_0$  of  $c_0$ ,  $c$  has at most one lift within distance  $2 \log p + 2\delta \log p$  of  $z_0$  by part (a). The claim then follows from the following lemma:

**Lemma 14.** *For all  $R > 0$  there exists  $M = R + O(1)$  such that the following is true: Let  $z, z' \in \mathbb{H}$  such that  $d(z, z') = 2D$ . Then  $B(z, D + R) \cap B(z', D + R)$  is contained in a ball of radius  $M$  centered at the midpoint of  $z$  and  $z'$ .*

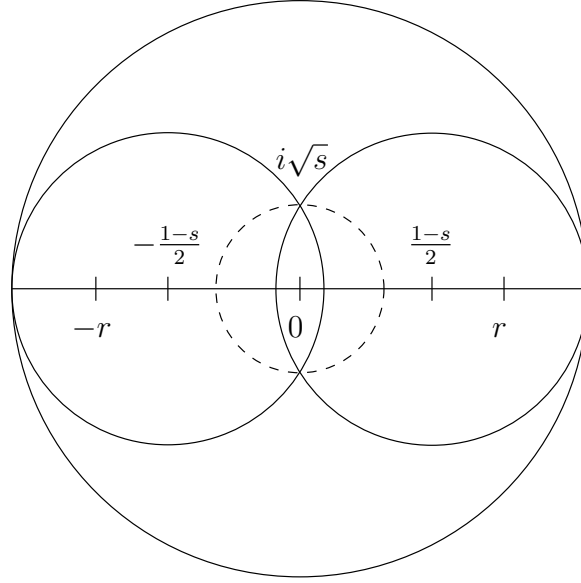


FIGURE 2.

*Proof.* We work in the Poincaré disk model. It suffices to consider  $z = r, z' = -r$  for an appropriate real point  $0 < r < 1$ . Let  $s = \tanh(R/2)$  so that  $s$  is distance  $R$  to 0. Then  $B(z, D + R)$  is contained in the euclidean disk  $B_e(\frac{1-s}{2}, \frac{1+s}{2})$  and  $B(z', D + R)$  is contained in the euclidean disk  $B_e(\frac{s-1}{2}, \frac{1+s}{2})$ . These 2 disks intersect in convex figure which is contained in the euclidean disk  $B_e(0, \sqrt{s})$  (see Figure 2), or equivalently the hyperbolic disk  $B(0, 2 \tanh^{-1}(\sqrt{s}))$ .

Now,  $\tanh(x) = 1 - 2e^{-2x} + O(e^{-4x})$  for  $x > 0$  and  $\tanh^{-1}(1 - \delta) = -\ln \delta/2 + O(1)$  for  $\delta < 1/2$ . Thus

$$\begin{aligned} 2 \tanh^{-1}(\sqrt{s}) &= 2 \tanh^{-1}(\sqrt{1 - 2e^{-R} + O(e^{-2R})}) \\ &= 2 \tanh^{-1}(1 - e^{-R} + O(e^{-2R})) \\ &= 2(R/2 + O(1)) \\ &= R + O(1) \end{aligned}$$

□

To prove (c), we note that by the proof of (a),  $x$  must be within  $\delta \log p + O(1)$  of a point mapping to  $c_2$ . Thus, it suffices to prove that  $i$  is within  $(1 + 2\delta) \log p + O(1)$  of at most  $O(p^{12\delta \log p})$  cusps. If  $c$  is such a cusp, then  $i$  must be within  $2\delta \log p + O(1)$  of a point  $x$  mapping to  $c_2$  neighboring  $c$ . By Lemma 11 there are at most  $O(e^{12\delta \log p}) = O(p^{12\delta})$  such points  $x$ . This completes the proof. □

## 4.2. Repulsion of Heegner CM points.

We endow the product  $X(p) \times X(p)$  with the product metric with respect to  $h_{X(p)}$  on each factor, so that for two points  $\xi = (x, y)$  and  $\xi' = (x', y')$  in  $X(p) \times X(p)$ ,

$$d_{X(p) \times X(p)}(\xi, \xi') = \max(d_{X(p)}(x, x'), d_{X(p)}(y, y'))$$

We'll also refer to  $d_{X(p) \times X(p)}$  simply as  $d$ . Note that  $d$  is equivalently the Kobayashi metric of  $X(p) \times X(p)$  (up to a constant), and that a ball in this metric is a product of balls on each factor:

$$B(\xi, R) = B(x, R) \times B(y, R)$$

**Proposition 15.** *For all sufficiently small  $\delta > 0$  and all sufficiently large  $p$ , if  $\xi, \xi' \in \text{CM}^+$  are distinct Heegner CM points in  $X(p) \times X(p)$  with the same projections to  $X(1)_p \times X(1)_p$  and  $B(\xi, \delta \log p) \cap B(\xi', \delta \log p) \neq \emptyset$ , then both  $\xi$  and  $\xi'$  lie on some Hecke divisor  $T_m$  with  $m = p^{O(\delta)}$ .*

*Proof.* Let  $\xi = (x, y)$  and  $\xi' = (x', y')$ . Diagonally acting by  $G(p)$  does not affect the statement of the proposition. We prove the case where  $x = \iota$ , as the case when  $x$  projects to  $c_3$  is similar.

Set  $x = gy$ , for  $g \in G(p)$ . Let  $t \in G(p)$  be a stabilizer of  $x$ , and take  $t_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in \text{SL}_2(\mathbb{Z})$  a lift of  $t$ , with  $\tilde{t} \in \text{GL}_2 \mathbb{F}_p$  its reduction. Since  $y$  must also be stabilized by  $t$ ,  $gtg^{-1}$  is either  $t$  or  $t^{-1}$  in  $G(p)$ . Moreover, by looking at the eigenvalues we see that  $\tilde{g}\tilde{t}\tilde{g}^{-1}$  is either  $\tilde{t}$  or  $\tilde{t}^{-1}$ . Since  $(x, y)$  is a Heegner CM point by assumption, it follows that (any lift of)  $g$  commutes with  $\tilde{t}$ .

Since  $d(\xi, \xi') < 2\delta \log p$ , it follows from Lemma 11 that there are  $M_x, M_y \in \text{SL}_2(\mathbb{Z})$  with  $h(M_x) = O(p^{4\delta})$  and  $h(M_y) = O(p^{4\delta})$  such that  $\gamma_p(M_x) \cdot x = x'$  and  $g^{-1}\gamma_p(M_y) \cdot x = y'$ . Then  $g^{-1}\gamma_p(M_y M_x^{-1})$  maps  $x'$  to  $y'$ , and so by the above it commutes with  $\gamma_p(M_x)\tilde{t}\gamma_p(M_x)^{-1}$ . Equivalently,  $h_y^{-1}gh_x$  commutes with  $\tilde{t}$ , where  $h_x = \gamma_p(M_x)$  and  $h_y = \gamma_p(M_y)$ .

Using the identification  $G(p) \cong \mathrm{PSL}_2(\mathbb{F}_p)$ , we equivalently have the two relations  $[t_0, g] = \mathbf{1}$  and  $[t_0, M_y^{-1}gM_x] = \mathbf{1}$ , which we view as a set of linear equations (defined over  $\mathbb{Z}$ ) in the coefficients of a  $2 \times 2$  matrix  $g \in M_2(\mathbb{F}_p)$  over  $\mathbb{F}_p$ .

**Lemma 16.** *For large enough  $p$ , the set of solutions  $g \in M_2(\mathbb{F}_p)$  to  $[t, g] = \mathbf{1}$  and  $[t, M_y^{-1}gM_x] = \mathbf{1}$  is at most 1-dimensional.*

*Proof.* Note that these linear equations have coefficients of size  $p^{O(\delta)}$ . The relation  $g\tilde{t} = \tilde{t}g$  has a 2-dimensional set of solutions in  $g$ , equal to the span  $\langle 1, \tilde{t} \rangle$  of the identity matrix and  $\tilde{t}$ . Thus, either the two relations define a line, or the second relation is redundant, and the second case is equivalent to

$$\langle M_y^{-1}M_x, M_y^{-1}\tilde{t}M_x \rangle \subset \langle 1, \tilde{t} \rangle.$$

Note that this holds if a set of minors vanishes over  $\mathbb{F}_p$ , and each of those minors is of size  $p^{O(\delta)}$ . Thus, for sufficiently small  $\delta$ , the second relation is redundant over  $\mathbb{F}_p$  if and only if it is redundant over  $\mathbb{Q}$ .

If the second relation is redundant over  $\mathbb{Q}$ , setting  $H$  to be the centralizer of  $t_0$ , we must have  $M_x^{-1}HM_y = H$ , which implies

$$M_x^{-1}HM_x = (M_x^{-1}HM_y)(M_x^{-1}HM_y)^{-1} = HH^{-1} = H.$$

Likewise,  $M_y^{-1}HM_y = H$ . Now, we claim that the elements of the normalizer of  $H$  in  $\mathrm{GL}_2(\mathbb{Q})$  which have positive norm consist exactly of  $H$ . To prove this, note that its enough to check it after tensoring with  $\mathbb{R}$ , in which case  $H$  becomes an embedded  $\mathbb{C}^* \subset \mathrm{GL}_2(\mathbb{R})$  (unique up to conjugation). Thus, as  $M_x, M_y$  have norm 1, we conclude that  $M_x, M_y \in H$ . Finally, note that since  $H \cap M_2(\mathbb{Z})$  is isomorphic  $\mathbb{Z}[i]$  we must have that both  $\gamma_p(M_x), \gamma_p(M_y)$  are stabilizers of  $x$ , contradicting the assumption that our Heegner CM points were distinct.  $\square$

Thus, the two relations must not be redundant, and we end up with a single projective solution  $g$ , which must lie in the span of  $\mathbf{1}, \tilde{t}$ . Now, as finding a kernel of a linear map is polynomial in the entries of a map, we can find an integral representative for  $g$  with entries of size  $p^{O(\delta)}$ . Thus  $g = a + b \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$  where  $\max(a, b) = O(p^\delta)$ , and  $(x, y)$  lies on  $T_m$  where  $m = \det g = a^2 + b^2$ .  $\square$

#### 4.3. Repulsion of singular bicusps.

**Proposition 17.** *For all sufficiently small  $\delta > 0$  and all sufficiently large  $p$ , if  $\xi \in X(p) \times X(p)$  is a point that is within  $(1 + \delta) \log p$  of at least 3 singular bicusps, then all the singular bicusps within  $(1 + \delta) \log p$  of  $\xi$  lie on the same Hecke divisor  $T_m$  with  $m = p^{O(\delta)}$ .*

*Proof.* Suppose  $\xi = (x, y)$  is a point as in the proposition. We first claim that each of  $x, y$  is within  $(1 + \delta) \log p$  of at least 2 distinct cusps. Suppose not, so that without loss of generality there exist a unique cusp  $c_0$  within  $(1 + \delta) \log p$  of  $x$ . That means that  $y$  is within  $(1 + \delta) \log p$  of at least 2 distinct cusps  $c, c'$  which have the same stabilizer as  $c_0$ . Without loss of generality, by acting with  $G(p)$  we can assume that  $c = iy_p$ . Thus, denoting  $H = \langle \sigma_p \rangle$ , by Lemma 12 we have elements  $\sigma \in H, M \in \mathrm{SL}_2 \mathbb{Z}$  such that  $h(M) = p^{O(\delta)}$  and  $c' = \sigma \gamma_p(M) \cdot c$ . Therefore  $\sigma \gamma_p(M) \in N(H)$ , and so  $\gamma_p(M) \in N(H)$ . Since  $h(M) = p^{O(\delta)}$  and the



only upper triangular matrices in  $\mathrm{SL}_2 \mathbb{Z}$  are strictly upper triangular, it follows that  $\gamma_p(M) \in H$ . Thus  $c = c'$ , which is a contradiction.

Thus each of  $x, y$  is within  $(1 + \delta) \log p$  of at least 2 distinct cusps, and also within  $3\delta \log p + O(1)$  of a pre-image of  $c_2$ . At the cost of decreasing  $\delta$  by a factor of 3 and acting by  $G(p)$  we can assume that  $x = \iota$  and  $y = gx$  for some  $g \in G(p)$ . Now suppose  $(c, c')$  is a singular bicuspid that is within  $(1 + \delta) \log p$  of  $(x, y)$ . Then it follows similarly to Lemmas 12 and 11 that there exist elements  $M_1, M_2 \in \mathrm{SL}_2(\mathbb{Z})$  with  $h(M_1), h(M_2) = p^{O(\delta)}$  such that  $c = \gamma_p(M_1) \cdot iy_p$  and  $c' = g\gamma_p(M_2) \cdot iy_p$ . It thus follows that

$$\gamma_p(M_1)^{-1}g\gamma_p(M_2) \in N(H),$$

or alternatively that

$$g \in \gamma_p(M_1)N(H)\gamma_p(M_2)^{-1}.$$

Now, note that  $G(p)$  acts on  $\mathbb{P}^1(\mathbb{F}_p)$  and  $\gamma_1 N(H) \gamma_2^{-1}$  are exactly those elements of  $G(p)$  that take  $\gamma_2 \infty$  to  $\gamma_1 \infty$  for any  $\gamma_1, \gamma_2 \in G(p)$ . Thus the intersection of any finite number of double cosets of the form  $\gamma_1 N(H) \gamma_2^{-1}$  is given by specifying the images of finitely many points in  $\mathbb{P}^1(\mathbb{F}_p)$ . Let  $A(g)$  denote the set of pairs  $(M_1, M_2)$  of matrices  $M_1, M_2 \in \mathrm{SL}_2(\mathbb{Z})$  for which  $g \in \gamma_p(M_1)N(H)\gamma_p(M_2)^{-1}$  and consider the intersection

$$B(g) := \bigcap_{(M_1, M_2) \in A(g)} \gamma_p(M_1)N(H)\gamma_p(M_2)^{-1}.$$

There are 2 cases:

- (1)  $B(g)$  specifies where at most 2 distinct points go, but no more. This means that for all the  $M_2$  that occur in  $A(g)$ ,  $\gamma_p(M_2)$  lies in at most 2 right  $N(H)$  orbits. However, since  $h(M_2) = p^{O(\delta)}$ , it follows that for large enough  $p$  there are at most 2 such  $\gamma_p(M_2)$ , which is a contradiction.
- (2)  $B(g)$  specifies where 3 distinct points go, and thus specifies  $g$ . Note that this means that there is a representative for  $g$  which is polynomial in the coefficients of the linear equations mod  $p$ , all of whose coefficients are reductions of elements in  $\mathbb{Z}$  of size  $p^{O(\delta)}$ . Thus there is an integral representative for  $g$  with entries of size  $p^{O(\delta)}$ . Since  $c = \gamma_p(M_1^{-1})g\gamma_p(M_2)$  it follows that  $(c, c')$  is on some  $T_m$  with  $m = \det g = p^{O(\delta)}$ , as desired.  $\square$

#### 4.4. Repulsion of diagonals.

Let  $\pi_i : (X(p) \times X(p))^2 \rightarrow X(p) \times X(p)$  be the projection onto the  $i$ th factor; we denote a point  $\xi \in (X(p) \times X(p))^2$  by  $\xi = (x_1, y_1, x_2, y_2)$  where  $\pi_i(\xi) = (x_i, y_i)$  for  $i = 1, 2$ . By the big diagonals of  $(X(p) \times X(p))^2$  we mean the subvarieties of the form

$$\Delta_g = \{(gx, gy, x, y)\} \subset (X(p) \times X(p))^2$$

for some fixed  $g \in G(p)$ , and by a small Hecke curve of degree  $m$  we'll mean

$$\tau_{g,m} = \{(gx, gy, x, y) \mid (x, y) \in T_m\} \subset \Delta_g$$

In the following proposition we'll be concerned with the Kobayashi neighborhoods of the big diagonals of the form

$$B(\Delta_g, R) = \{(x_1, y_1, x_2, y_2) \mid d(x_1, gx_2) < R \text{ and } d(y_1, gy_2) < R\}$$

**Proposition 18.** *For all sufficiently small  $\delta > 0$ , and all sufficiently large  $p$ , if a point  $\xi \in (X(p) \times X(p))^2$  is in  $\omega(\log p)$  many distinct neighborhoods  $B(\Delta_g, \delta \log p)$  then one of the following must be true:*

- (a)  $\xi$  is within  $(1 + O(\delta)) \log p$  of a point both of whose projections are singular bicusps. In this case,  $\xi$  is in  $O(p^{1+\delta/2}e^{-d})$  many such neighborhood where  $d$  is the smaller of the distances of  $\pi_1(\xi)$  and  $\pi_2(\xi)$  to a singular bicusps;
- (b)  $\xi$  is within  $O(\delta \log p)$  of a small Hecke curve  $\tau_{g,m}$  of degree  $m = p^{O(\delta)}$ .

*Proof.* Let  $\xi = (x_1, y_1, x_2, y_2)$  be a point in  $\omega(\log p)$  many neighborhoods  $B(\Delta_g, \delta \log p)$ . We split the proof up into 2 cases:

Case 1: First suppose one of the coordinates, say  $x_1$ , is within  $(1 - 2\delta) \log p$  of a cusp  $c$ , which after acting by an element of  $G(p)$  we may assume to be the image of  $iy_p$ . For each  $g \in G(p)$  such that  $d(x_1, gx_2) < \delta \log p$  it must be the case that the cusp nearest to  $gx_2$  is also  $c$ , by Lemma 5. Hence, since  $\xi$  is in many diagonal neighborhoods,  $g$  must be in a left  $H$  coset, where  $H$  is the stabilizer of  $c$ . Without loss of generality, we can similarly assume the nearest cusp to  $x_2$  is  $c$  and therefore the set of all  $g$  such that  $\xi \in B(\Delta_g, \delta \log p)$  is inside  $H$ .

Next, let  $c' = g_0c$  be the cusp closest to  $y_1$ , and assume that  $y_1$  is not within  $(1 + \delta) \log p$  of any cusp stabilized by  $H$ . Then  $y_1$  can't be within  $(1 - \delta) \log p$  of  $c'$ , or else there would be no  $h \in H$  such that  $d(y_1, hy_1) < 2\delta \log p$ , which must be the case since  $\xi \in B(\Delta_h, \delta \log p) \cap B(\Delta_1, \delta \log p)$ . Thus  $y_1$  is within  $\delta \log p$  of a point projecting to  $c_2$ , which we may assume to be  $g_0\iota$ . Now, for each  $h \in H$ ,  $d(g_0\iota, hg_0\iota) < 3\delta \log p$ , so by Lemma 11 there is a matrix  $M \in \text{SL}_2(\mathbb{Z})$  with  $h(M) = O(p^{6\delta})$  such that  $\gamma_p(M) = g_0^{-1}hg_0$ .  $M$  must be unipotent, so  $g_0 \in N(H)\gamma_p(M')$  for some  $M' \in \text{SL}_2(\mathbb{Z})$  with  $h(M') = p^{O(\delta)}$ , by the following lemma.

**Lemma 19.** *Let  $M_1, M_2 \in \text{SL}_2(\mathbb{Z})$  be unipotent with  $h(M_1), h(M_2) = O(p^\delta)$ , and  $A \in \text{PGL}_2 \mathbb{F}_p$  such that  $M_1$  and  $AM_2A^{-1}$  generate the same unipotent subgroup mod  $p$ . Then  $A \in N(\langle \gamma_p(M_2) \rangle) \gamma_p(B) \bmod p$  where  $B \in \text{SL}_2(\mathbb{Z})$  has height  $p^{O(\delta)}$ .*

*Proof.*  $M_1, M_2$  fix unique vectors  $u, v \in \mathbb{Z}^2$  (up to scale) whose components are of size  $p^{O(\delta)}$ . There is then  $B \in \text{SL}_2(\mathbb{Z})$  with  $h(B) = p^{O(\delta)}$  taking  $u$  to  $v$ , and  $AB^{-1}$  fixes  $v \bmod p$ . Thus  $AB^{-1} \in N(\langle M_2 \rangle) \bmod p$ . The claim follows.  $\square$

Thus, this means that  $g_0\iota$  is within  $\log p + O(\delta \log p)$  of a cusp stabilized by  $H$ . Therefore  $y_1$  is as well, and  $\pi_1(\xi)$  is within a distance  $M \leq (1 + O(\delta)) \log p$  of a singular bicusps; by the same argument,  $\pi_2(\xi)$  is also.

Finally, we can calculate the distance between  $\xi$  and  $\sigma_p^k \xi$  in each projection (to  $X(p)$ ) in the disk surrounding  $c$ , since the injectivity radius is close to  $2 \log p$ . We thus have use of the following lemma

**Lemma 20.** *For  $R = 1 - \epsilon$ ,  $\theta \in \mathbb{R}$  such that  $\epsilon = o(1)$ ,  $\theta = o(1)$  and  $\epsilon = o(\theta)$  we have*

$$\left| \frac{R - Re^{i\theta}}{1 - R^2 e^{i\theta}} \right| = 1 - \frac{2\epsilon^2}{\theta^2} (1 + o(1))$$

*Proof.* Consider  $\theta$  fixed and write  $F(R) = R|1 - e^{i\theta}|^2$  and  $G(R) = |1 - R^2 e^{i\theta}|^2$ .  $F(1) = G(1) = 4 \sin^2 \frac{\theta}{2}$ . Next note that

$$F(R) = R^2(1 - 2 \cos \theta + \cos^2 \theta + \sin^2 \theta) = R^2(2 - 2 \cos \theta) = 4R^2 \sin^2 \frac{\theta}{2}$$

So that  $F'(R) = 8R \sin^2 \frac{\theta}{2}$  and  $F''(R) = 8 \sin^2 \frac{\theta}{2}$ .

Likewise,

$$G'(R) = 2\Re(-2Re^{i\theta}(1 - R^2 e^{-i\theta})) = 4R^3 - 4R \cos \theta$$

and  $G''(R) = 12R - 4 \cos \theta$ . Thus  $G'(1) = 4 - 4 \cos \theta = 8 \sin^2 \frac{\theta}{2}$ . Thus, since  $\epsilon = o(\theta)$  by Taylors theorem we have

$$\left| \frac{R - Re^{i\theta}}{1 - R^2 e^{i\theta}} \right|^2 = \frac{4 \sin^2 \frac{\theta}{2} (1 - 2\epsilon + \epsilon^2)}{4 \sin^2 \frac{\theta}{2} (1 - 2\epsilon) + \epsilon^2 (6 - 2 \cos \theta) + O(\epsilon^3)} = 1 - \frac{4\epsilon^2}{\theta^2} (1 + o(1))$$

which implies the result.  $\square$

Hence, if  $kp^{-1} = \omega(e^{-M})$  then we have

$$\tanh(d(\xi, \sigma_p^k \xi)/2) = 1 - 8e^{-2M} p^2 k^{-2} (1 + o(1)).$$

Thus, if  $d(\xi, \sigma_p^k \xi) < \delta \log p$  we must have

$$k^{-1} p e^{-M} \gg p^{-\delta/2},$$

or  $k \ll p^{1+\delta/2} e^{-M}$ .

Case 2: Now assume that none of the coordinates of  $\xi$  are within  $(1 - 2\delta) \log p$  of a cusp, and we show that  $\xi$  must be within  $O(\delta \log p)$  of a small Hecke curve. We can assume  $\xi = (\gamma_1 \iota, \gamma_2 \iota, \gamma_3 \iota, \gamma_4 \iota)$  as  $\xi$  is within a radius of  $O(\delta \log p)$  of such a point. For each  $g$  such that  $\xi \in B(\Delta_g, \delta \log p)$ , by Lemma 11 there exist  $M_g, M'_g \in \text{SL}_2(\mathbb{Z})$  with  $h(M_g), h(M'_g) = O(p^{3\delta})$  such that  $\gamma_p(M_g) = \gamma_1^{-1} g \gamma_3$  and  $\gamma_p(M'_g) = \gamma_2^{-1} g \gamma_4$ , or in other words that

$$\gamma_1 \gamma_p(M_g) \gamma_3^{-1} = \gamma_2 \gamma_p(M'_g) \gamma_4^{-1}$$

Two distinct such elements  $g, h$  would then yield matrices  $N = M_g M_h^{-1}$  and  $N' = M'_g M'_h^{-1}$  with  $h(N), h(N') = O(p^{6\delta})$  such that

$$\gamma_3 \gamma_p(N) \gamma_3^{-1} = \gamma_4 \gamma_p(N') \gamma_4^{-1}$$

or equivalently

$$\gamma_p(N) = \gamma \gamma_p(N') \gamma^{-1}$$

for  $\gamma = \gamma_3^{-1} \gamma_4$ .

Defining

$$S_\delta := \{M \in \text{SL}_2(\mathbb{Z}) \mid h(M) = O(p^{6\delta})\},$$

and  $\overline{S}_\delta \subset \text{PSL}_2(\mathbb{F}_p)$  its reduction mod  $p$ , we see that the number of diagonal neighborhoods containing  $\xi$  is bounded by

$$|\gamma_p(\overline{S}_\delta) \cap \gamma \gamma_p(\overline{S}_\delta) \gamma^{-1}|.$$

Now consider the larger set  $S'_\delta = \{M \in M_2(\mathbb{Z}) \mid h(M) = O(p^{6\delta})\}$ ,  $\overline{S}'_\delta \subset M_2(\mathbb{F}_p)$  its reduction, and the subspace

$$T := \text{Span}(\gamma_p(\overline{S}'_\delta) \cap \gamma \gamma_p(\overline{S}'_\delta) \gamma^{-1}) \subset M_2(\mathbb{F}_p).$$

We now separate into two cases:

- (1) The centralizer of  $T$  consists of more than just scalars. It follows that  $T$  is a sub-algebra, and so it must either be a torus, or isomorphic to  $\mathbb{F}_p[x]/(x^2)$ . If  $T \cong \mathbb{F}_p[x]/(x^2)$  then all the elements in

$$\gamma_p(\overline{S}_\delta) \cap \gamma \gamma_p(\overline{S}_\delta) \gamma^{-1}$$

are in a single unipotent subgroup  $U$ . Thus as in the analysis of  $y_1$  in case 1 all the co-ordinates of  $\xi$  must be within  $(1 + O(\delta)) \log p$  of a cusp stabilized by  $U$  and we are in case (a) of the proposition.

If  $T$  is a torus, by picking a non-scalar element  $\gamma_p(M) \in T$  we can lift  $T$  to a torus  $\tilde{T} \subset M_2(\mathbb{Q})$  spanned by  $\mathbf{1}$  and  $M$ . Thus the elements in

$$\gamma_p(\overline{S}_\delta) \cap \gamma \gamma_p(\overline{S}_\delta) \gamma^{-1}$$

are reductions of elements in the norm 1 subgroup of  $\tilde{T}$ , and hence are generated by a single semisimple element  $\gamma_p(M)$ . As there are at most  $O(\log p)$  elements  $M^k$  with height bounded by  $p^{O(\delta)}$ , it cannot be the case that  $\xi$  is in  $\omega(\log p)$  many diagonal neighborhoods.

- (2) The centralizer of  $T$  consists of scalars. This means we can pick at most three elements in  $A_1, A_2, A_3 \in T$  such that they have no common centralizer outside of scalars. Thus,  $\gamma$  is determined projectively by the three elements  $\gamma A_i \gamma^{-1}$ . Since  $h(A_i) = O(p^{6\delta})$ , this means we can find a projective representative  $\tilde{\gamma} \in \text{GL}_2(\mathbb{F}_p)$  for  $\gamma$  with entries of size  $p^{O(\delta)}$ . Thus, by gaussian elimination, we can find elements  $M_1, M_2 \in \text{SL}_2(\mathbb{Z})$  of height  $p^{O(\delta)}$  such that

$$\gamma_p(\gamma'_1) \tilde{\gamma} \gamma_p(\gamma'_2) = \begin{pmatrix} 0 & m \\ -1 & 0 \end{pmatrix},$$

where  $m = \det \tilde{\gamma} = p^{O(\delta)}$ .

Next, note that  $d_{Y(p)}(i, i\sqrt{m}) = \frac{1}{2} \log m$ , and that

$$\left( i\sqrt{m}, \begin{pmatrix} 0 & m \\ -1 & 0 \end{pmatrix} \cdot i\sqrt{m} \right) \in T_m.$$

Thus, since the metric  $h_{Y(p)}$  on  $Y(p)$  is strictly smaller than the metric  $h_{X(p)}$ , by the above and Lemma 7 we have

$$\begin{aligned} d_{X(p)}((\gamma_3 \iota, \gamma_4 \iota), T_m) &= d_{X(p)}((\iota, \gamma \iota), T_m) \\ &\leq d_{X(p)}(\gamma_p(M_1^{-1}) \iota, \iota) + d_{X(p)}(\gamma_p(M_2) \iota, \iota) + d_{X(p)}((\gamma_p(M_1) \iota, \gamma \gamma_p(M_2) \iota), T_m) \\ &\leq O(\delta \log p) + d_{Y(p)}((\iota, \begin{pmatrix} 0 & m \\ -1 & 0 \end{pmatrix} \iota), T_m) \\ &\leq O(\delta \log p) + 2d_{Y(p)}(i, i\sqrt{m}) \\ &\leq O(\delta \log p) \end{aligned}$$

Which establishes the claim, since  $m = p^{O(\delta)}$ .

□

## 5. VOLUME ESTIMATES

In this section we prove that, for certain special subvarieties  $V$  of hyperbolic manifolds, the total volume of a curve in a neighborhood of  $V$  of radius  $r$  grows sharply as a function of  $r$ . This has two consequences: one can effectively bound the volume in such a neighborhood by the volume in a larger neighborhood, and in the limit  $r \rightarrow 0$  one can effectively bound the multiplicity of a curve along such a subvariety by the volume in a neighborhood of it. In both cases, bigger neighborhoods give better bounds.

## 5.1. Global volume estimates.

Let  $X$  now be a hyperbolic curve and consider any curve  $C \subset X \times X$ . Work of Hwang and To [HT02, HT12] provides a bound on the multiplicity of  $C$  at a point  $\xi \in X \times X$  in terms of the volume of  $C$  in a (Kobayashi) ball centered at  $x$ . Similarly, the multiplicity of  $C$  along the diagonal  $\Delta \subset X \times X$  is bounded by its volume within the tubular neighborhood

$$B(\Delta, r) = \{(x, y) \mid d(x, y) < r\} \subset X \times X$$

of the diagonal. Note that this is the Kobayashi tube of radius  $r/2$  around  $\Delta$ . For  $r < \rho_X$ ,  $B(\Delta, r)$  is the quotient of the neighborhood

$$B(\Delta_{\mathbb{D}}, r) = \{(z, w) \mid d(z, w) < r\} \subset \mathbb{D} \times \mathbb{D}$$

by the diagonal action of  $\pi_1(X)$ .

Throughout the following we equip  $X \times X$  with its product metric as in Section 4.2.  $X \times X$  is naturally endowed with a Kähler form  $\omega_{std}$  which is the sum of the pullbacks of the Kähler form associated to  $h_X$  along each projection. Equivalently,  $\omega_{std}$  descends from the form

$$\omega_{std} = \frac{2idz \wedge d\bar{z}}{(1 - |z|^2)^2} + \frac{2idw \wedge d\bar{w}}{(1 - |w|^2)^2}$$

on  $\mathbb{D} \times \mathbb{D}$ . Volumes in  $X \times X$  are taken with respect to this form. We then have:

**Theorem 21.** *For any curve  $C \subset X \times X$ :*

- (a) [HT02, Theorem 2] *For any point  $\xi \in X \times X$ , and  $r < \rho_X$ , then*

$$\text{vol}(C \cap B(\xi, r)) \geq 4\pi \sinh^2(r/2) \text{mult}_{\xi}(C)$$

- (b) [HT12, Theorem 1] *For any  $r < \rho_X$ ,*

$$\text{vol}(C \cap B(\Delta, r)) \geq 8\pi \sinh^2(r/4)(C \cdot \Delta)$$

Both statements in Theorem 21 are optimal in the sense that the bound is realized by a union of translates of the image of the graph of  $-z : \mathbb{D} \rightarrow \mathbb{D}$ . For the convenience of the reader, we summarize the proof of part (b) above. Recall the following

**Definition.** For  $\varphi(z)$  a plurisubharmonic function on a neighborhood of a point  $x$  in some complex manifold  $M$ , the Lelong number of  $\varphi$  at  $x$  is

$$\nu(\varphi, x) := \liminf_{z \rightarrow x} \frac{\varphi(z)}{\log |z - x|}.$$

For example, if  $V \subset M$  is a divisor cut out locally by  $f$ , and  $\varphi(z) = \log |f(z)|$ , then  $\nu(\varphi, x) = \text{mult}_x(f)$ .

Hwang and To define a plurisubharmonic function  $F : \mathbb{D} \times \mathbb{D} \rightarrow \mathbb{R}$  that is diagonally-invariant under the full isometry group  $\text{PSL}_2 \mathbb{R}$  such that  $0 \leq \omega_F = i\partial\bar{\partial}F \leq \omega_{std}$ , as well as diagonally-invariant functions  $f_\epsilon : \mathbb{D} \times \mathbb{D} \rightarrow \mathbb{R}$  that

- (1) are plurisubharmonic off the diagonal  $\Delta_{\mathbb{D}} \subset \mathbb{D} \times \mathbb{D}$ ;
- (2) agree with  $F$  outside of  $B(\Delta_{\mathbb{D}}, r)$ ;
- (3) have a logarithmic pole along the diagonal, and for any  $\xi \in \Delta_{\mathbb{D}}$ ,

$$\liminf_{\epsilon \rightarrow 0} \nu(f_\epsilon, \xi) = 8 \sinh^2(r/4)$$

As  $f_\epsilon$  and  $F$  descend to functions on  $X \times X$ , it then follows that for any curve  $C \subset X \times X$ ,

$$\text{vol}(C \cap B(\Delta, r)) \geq \int_{C \cap B(\Delta, r)} \omega_F = \int_{C \cap B(\Delta, r)} \omega_{f_\epsilon} \geq \pi \sum_{\xi \in C \cap \Delta} \nu(f_\epsilon, \xi) \text{mult}_\xi C$$

where we've used the diagonal invariance to descend the forms to  $X \times X$ . The equality follows from Stoke's theorem and the second inequality from the fact that, for  $[C]$  denoting the current of integration along  $C$ ,

$$\nu([C] \wedge \omega_{f_\epsilon}, \xi) \geq \nu([C], \xi) \nu(f_\epsilon, \xi)$$

(cf [HT02], Proposition 2.2.1(a)).

We shall require an analogue of the above theorem for the diagonal

$$\Delta_2 = \{(x, y, x, y)\} \subset (X \times X)^2.$$

Around  $\Delta_2$  we have the (Kobayashi) tubular neighborhood considered in the previous section

$$B(\Delta_2, r) = \{(x_1, y_1, x_2, y_2) \mid d(x_1, x_2) < r \text{ and } d(y_1, y_2) < r\} \subset (X \times X)^2$$

for any  $r < \rho_X$ , and it is the quotient of the analogous diagonal neighborhood  $B(\Delta_2, r) \subset (\mathbb{D} \times \mathbb{D})^2$  by the diagonal action of  $\pi_1(X)^2$ .

**Lemma 22.** *For  $X, r$  as in Theorem 21 and for any curve  $C \subset (X \times X)^2$  not contained in  $\Delta_2$ , we have*

$$\text{vol}(C \cap B(\Delta_2, r)) \geq 8\pi \sinh^2(r/4) \sum_{\xi \in \Delta_2} \text{mult}_\xi C$$

*Proof.* Let  $\pi_i : (X \times X)^2 \rightarrow X \times X$  be the two projections for  $i = 1, 2$ , and let  $F_i = \pi_i^* F$ ,  $f_{\epsilon, i} = \pi_i^* f_\epsilon$ ,  $G = \max(F_1, F_2)$  and  $g_\epsilon = \max(f_{\epsilon, 1}, f_{\epsilon, 2})$ . Then  $G$  and  $g_\epsilon$  obey the same properties with respect to  $B(\Delta_2, r)$  as  $F$  and  $f_\epsilon$  in the discussion above. Specifically,  $0 \leq \omega_G \leq \omega_{std}$ , and  $\omega_{g_\epsilon}$  is positive off of the diagonal and equal to  $\omega_G$  outside of  $B(\Delta_2, r)$ . The Lelong number of  $\omega_{g_\epsilon}$  along the diagonal is the minimum of the those of  $\omega_{g_1}$  and  $\omega_{g_2}$ , and therefore we again have

$$\text{vol}(C \cap B(\Delta_2, r)) \geq \int_{C \cap B(\Delta_2, r)} \omega_G = \int_{C \cap B(\Delta_2, r)} \omega_{g_\epsilon} \geq \pi \sum_{\xi \in C \cap \Delta_2} \nu(g_\epsilon, \xi) \text{mult}_\xi C$$

The claim follows.  $\square$

*Remark 23.* Using the above strategy and the fact that the Kobayashi metric can be written (*cf.* [HT02, Section 3]) as the maximum of metrics pulled back from  $\mathbb{D}$ , it is possible to prove an analogue of Lemma 22 for any diagonal in powers of symmetric domains.

## 5.2. Relative volume estimates.

The result of Section 5 describes how the volume of a curve  $C \subset X \times X$  contained within a tube around the diagonal  $\Delta \subset X \times X$  grows as a function of the radius  $r$ . In this section, we prove a similar result comparing the volume within a radius  $R$  to that within a smaller radius  $r$ .

**Proposition 24.** *For  $X$  a compact hyperbolic complex curve, any complex curve  $C \subset X \times X$  and any  $\rho_X > R > r > 0$ ,*

$$\text{vol}(C \cap B(\Delta, R)) \geq \frac{\cosh(R/2)}{\cosh(r/2)} \text{vol}(C \cap B(\Delta, r))$$

*Remark 25.* Note that the constant in Proposition 24 is presumably not optimal.

*Proof.* The proof uses many of the techniques of [HT12]. In the Poincaré disk model, take

$$\chi(z, w) = \tanh^2(d_{\mathbb{D}}(z, w)/2) = \left| \frac{w - z}{1 - \bar{z}w} \right|^2$$

and consider the potential  $F(z, w) = f(s)$  on  $\mathbb{D}$ , where  $s = \log \chi(z, w)$ . Let  $C = \log \tanh^2(R/2)$  and  $c = \log \tanh^2(r/2)$ .  $F$  is invariant under the diagonal action of  $\text{SL}_2 \mathbb{R}$ , so the form  $i\partial\bar{\partial}F$  is determined by its value at  $(0, w)$ , where

$$\begin{aligned} \omega_F &:= i\partial\bar{\partial}F \\ &= f'(s) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + f''(s)e^{-s} \begin{pmatrix} (1 - e^s)^2 & e^s - 1 \\ e^s - 1 & 1 \end{pmatrix} \end{aligned}$$

The matrix coefficients represent the coefficients in the obvious basis of (1,1)-forms, according to

$$\begin{pmatrix} idz \wedge d\bar{z} & idz \wedge d\bar{w} \\ idw \wedge d\bar{z} & idw \wedge d\bar{w} \end{pmatrix}$$

Note that for  $f_0(s) = -2\log(e^{-s} - 1)$  and  $F_0(z, w) = f_0(s)$ , the off-diagonal terms cancel, and we have (at  $(0, w)$ )

$$\omega_{F_0} = 2 \begin{pmatrix} 1 & 0 \\ 0 & (1 - |w|^2)^{-2} \end{pmatrix} = \omega_{std}$$

on  $\mathbb{D} \times \mathbb{D} - \Delta_{\mathbb{D}}$ , since  $\chi(0, w) = |w|^2$ . Indeed as currents we have

$$\begin{aligned} \omega_{F_0} &= -2i\partial\bar{\partial} \log \left( \frac{1 - \chi(z, w)}{\chi(z, w)} \right) \\ &= -2i\partial\bar{\partial} \log(|1 - \bar{z}w|^2 - |z - w|^2) + 2i\partial\bar{\partial} \log |z - w|^2 \\ &= -2i\partial\bar{\partial} \log [(1 - |z|^2)(1 - |w|^2)] + 4\pi[\Delta_{\mathbb{D}}] \\ &= \omega_{std} + 4\pi[\Delta_{\mathbb{D}}] \end{aligned} \tag{2}$$

Take  $f(s)$  to be constant for  $s < c$ , equal to  $f_0$  for  $s > C$ , and satisfying  $f'(s) = \frac{2A}{\sqrt{1-e^s}} + \frac{2B}{1-e^s}$  for  $c \leq s \leq C$ , for  $A$  and  $B$  chosen so that  $f'(s)$  is continuous. Explicitly,



$$A = \frac{-\cosh(r/2)\cosh(R/2)}{\cosh(R/2) - \cosh(r/2)}, \quad B = \frac{\cosh(R/2)}{\cosh(R/2) - \cosh(r/2)}$$

We have  $f''(s) = \frac{Ae^s}{(1-e^s)^{3/2}} + \frac{2Be^s}{(1-e^s)^2}$ , so in the region  $c < s < C$ ,

$$\omega_F = B\omega_{std} + A \begin{pmatrix} (1-e^s)^{1/2} & (1-e^s)^{-1/2} \\ (1-e^s)^{-1/2} & (1-e^s)^{-3/2} \end{pmatrix}.$$

Note that the  $A$  coefficient is a positive semi-definite matrix, so that if  $A < 0$  we have the bound  $\omega_F \leq B\omega_{std}$ .

We thus have a function  $f = f(r, R)$ , whose corresponding  $\omega_F$  satisfies

- (1)  $\omega_F = 0$  for  $d(z, w) < r$ ;
- (2)  $\omega_F \leq B\omega_{std}$  for  $r < d(z, w) < R$ ;
- (3)  $\omega_F = \omega_{std}$  for  $d(z, w) > R$ .

$f(r, R)$  is not  $C^2$  at  $s = c$  and  $s = C$ , but replacing  $f$  by

$$g(s) = (2\epsilon)^{-1} \int_{|y| < \epsilon} f(r+y, R-y)(s) dy$$

for sufficiently small  $\epsilon > 0$ , it is easy to see that  $g(s)$  is plurisubharmonic, and forming  $G(z, w) = g(s)$  we have

- (1)  $\omega_G = 0$  for  $d(z, w) < r - \epsilon$ ;
- (2)  $\omega_G \leq (B + O(\epsilon))\omega_{std}$  for  $r - \epsilon < d(z, w) < R + \epsilon$ ;
- (3)  $\omega_G = \omega_{std}$  for  $d(z, w) > R + \epsilon$ .

By descending to  $X \times X$  we have by (2):

$$\begin{aligned} \text{vol}(C \cap B(\Delta, R)) &= \int_{C \cap B(\Delta, R)} \omega_{F_0} - 4\pi(C \cdot \Delta) \\ &= \int_{C \cap B(\Delta, R)} \omega_G - 4\pi(C \cdot \Delta) \\ &\leq (B + O(\epsilon)) \cdot (\text{vol}(C \cap B(\Delta, R)) - \text{vol}(C \cap B(\Delta, r))) \end{aligned}$$

where the second equality follows from Stoke's theorem (and the plurisubharmonicity of  $F_0$ , as in [HT02]). Letting  $\epsilon \rightarrow 0$  and simplifying, we get the result.  $\square$

A similar result holds for the growth of the volume of a curve near the *conjugate* diagonal. For a curve  $X$ , its conjugate  $\overline{X}$  is the same curve with the negated complex structure, and the pointwise diagonal  $\overline{\Delta} \subset X \times \overline{X}$  is called the conjugate diagonal.

**Proposition 26.** *For  $X$  a compact hyperbolic complex curve, any complex curve  $C \subset X \times \overline{X}$  that is not the conjugate diagonal, and any  $\rho_X > R > r > 0$ ,*

$$\text{vol}(C \cap B(\overline{\Delta}, R)) \geq \frac{\sinh(R/2)}{\sinh(r/2)} \text{vol}(C \cap B(\overline{\Delta}, r))$$

*Furthermore, the bound is optimal in the following sense: suppose  $X$  is isomorphic to  $\overline{X}$  via a map  $z \rightarrow \bar{z}$ . (For instance,  $X$  is defined over  $\mathbb{R}$ ). Then the graph  $(z, \bar{z})$  achieves the bound.*

*Proof.* The proof is very similar to Proposition 24. Suppose  $X = \Gamma \backslash \mathbb{D}$ , so that  $\bar{X} = \bar{\Gamma} \backslash \mathbb{D}$ , and consider the function  $\psi$  on  $\mathbb{D} \times \mathbb{D}$  given by

$$\psi(z, w) = \tanh^2(d_{\mathbb{D}}(z, \bar{w})/2) = \left| \frac{\bar{w} - z}{1 - zw} \right|^2.$$

$\psi$  is invariant under the diagonal action of  $\mathrm{SL}_2 \mathbb{R}$ . For any function  $f : \mathbb{R} \rightarrow \mathbb{R}$ , we compute that at  $(0, w)$ , the potential  $F(z, w) = f(\psi)$  yields a form

$$\begin{aligned} \omega_F &= i\partial\bar{\partial}F \\ &= f'(\psi) \begin{pmatrix} (1 - |w|^2)^2 & w^2 \\ \bar{w}^2 & 1 \end{pmatrix} + f''(\psi) \begin{pmatrix} |w|^2(1 - |w|^2)^2 & -w^2(1 - |w|^2) \\ -\bar{w}^2(1 - |w|^2) & |w|^2 \end{pmatrix} \end{aligned}$$

Taking  $s(\psi) = -\log(1 - \psi)$  and  $S(z, w) = s(\psi(z, w))$ , for instance, we have by direct computation

$$\omega_S = \begin{pmatrix} 1 & 0 \\ 0 & (1 - |w|^2)^{-2} \end{pmatrix} = \frac{\omega_{std}}{2}$$

where  $\omega_{std}$  is the standard form on  $\mathbb{D} \times \mathbb{D}$ . Let  $C = s(\tanh^2(R/2))$ ,  $c = s(\tanh^2(r/2))$ , and define a continuous function  $f : [0, 1] \rightarrow \mathbb{R}$  on the interval  $[\tanh^2(r/2), \tanh^2(R/2)]$  by  $f(\psi) = h(s(\psi))$ , where

$$h'(s) = \frac{1 - \sqrt{\frac{e^c - 1}{e^s - 1}}}{1 - \sqrt{\frac{e^c - 1}{e^C - 1}}}$$

Take  $f$  to be constant on  $[0, \tanh^2(r/2)]$ , and linear of slope 1 on  $[\tanh^2(R/2), 1]$ . One can easily compute that the resulting  $\omega_F$  is positive, and dominated by

$$(h'(s) + 2(1 - e^{-s})h''(s)) \frac{\omega_{std}}{2} = \left(1 - \sqrt{\frac{e^c - 1}{e^C - 1}}\right)^{-1} \frac{\omega_{std}}{2}$$

on (the interior of)  $B(\bar{\Delta}_{\mathbb{D}}, R) - B(\bar{\Delta}_{\mathbb{D}}, r)$ . Further we have that

$$\omega_F|_{B(\bar{\Delta}_{\mathbb{D}}, r)} = 0 \quad \text{and} \quad \omega_F|_{B(\bar{\Delta}_{\mathbb{D}}, \rho_X) - B(\bar{\Delta}_{\mathbb{D}}, R)} = \frac{\omega_{std}}{2}$$

Smoothing  $F$  out by the same trick as in the proof of Proposition 24 and descending these forms down to  $X \times \bar{X}$ , we have that

$$\begin{aligned} \mathrm{vol}(C \cap B(\bar{\Delta}, R)) &= \int_{C \cap B(\bar{\Delta}, R)} \omega_{std} \\ &= 2 \int_{C \cap B(\bar{\Delta}, R)} \omega_F \\ &\leq \left(1 - \sqrt{\frac{e^c - 1}{e^C - 1}}\right)^{-1} (\mathrm{vol}(C \cap B(\bar{\Delta}, R)) - \mathrm{vol}(C \cap B(\bar{\Delta}, r))) \end{aligned}$$

yielding the statement, as  $e^c - 1 = \sinh^2(r/2)$ , and likewise for  $C$  and  $R$ .  $\square$

## 6. MULTIPLICITY ESTIMATES

Consider the product  $(X(p) \times X(p))^n$ , and denote by  $\pi_i, \pi_{ij}$  the projections onto the  $i$ th and  $ij$ th factors, respectively. For the proof of Theorem 31, we will need to control the ramification of curves  $C \subset (X(p) \times X(p))^n$  over their image in  $\text{Sym}^n Z(p)$ . Such ramification occurs when  $C$  passes through one of the sets

$$\pi^{-1}(\text{CM}^+), \quad \pi_i^{-1}(\text{CM}^-), \quad \pi_i^{-1}(\text{SBC}), \quad \pi_{ij}^{-1}(\Delta_g),$$

for some  $i, j$ , where  $\Delta_g \subset (X(p) \times X(p))^2$  is the diagonal considered in Section 4.4.

In this section we prove that incidence of  $C$  along each of these sets is negligible with respect to its volume. For any set  $S$  of (closed) points, let  $\text{mult}_S(C) = \sum_{x \in S} \text{mult}_x(C)$ . Recall that by  $T_m^* C$  we mean the pullback of the divisor  $C$  along the Hecke correspondence in the second variable; clearly  $\text{vol}(T_m^* C) \leq \deg T_m \cdot \text{vol}(C)$ .

### 6.1. Multiplicity in $X(p) \times X(p)$ .

**Proposition 27.** *For all sufficiently small  $\delta > 0$ , all sufficiently large  $p$ , and for any non-Hecke curve  $C \subset X(p) \times X(p)$ ,*

$$\text{mult}_{\text{CM}^+}(C) = O(p^{-\delta} \text{vol}(C))$$

*Proof.* For  $d = p^\delta$ , partition  $\text{CM}^+$  into two sets

$$T := \text{CM}^+ \cap \bigcup_{m < d} T_m \quad \text{and} \quad S := \text{CM}^+ - T$$

By Proposition 15, for sufficiently small  $\delta' > 0$  the balls  $B(\xi, \delta' \log p)$  are disjoint as  $\xi$  varies over  $S$ . By Theorem 21 and Lemma 4 we then have that

$$\begin{aligned} \text{mult}_{\text{CM}^+}(C) &= \sum_{\xi \in S} \text{mult}_\xi(C) + \sum_{\xi \in T} \text{mult}_\xi(C) \\ &\ll \sinh(\delta' \log p / 2)^{-2} \text{vol}(C \cap \bigcup_{\xi \in S} B(\xi, \delta' \log p)) + \sum_{m < d} (C \cdot T_m) \\ &\ll p^{-\delta'} \text{vol}(C) + \sum_{m < d} (T_m^* C \cdot \Delta) \\ &\ll p^{-\delta'} \text{vol}(C) + \sinh(\log p / 2)^{-2} \sum_{m < d} (\deg T_m) \text{vol}(C) \\ &\ll (p^{-\delta'} + d^3 p^{-1}) \text{vol}(C) \end{aligned}$$

and the result follows.  $\square$

**Proposition 28.** *For all sufficiently small  $\delta > 0$ , all sufficiently large  $p$ , and for any non-Hecke curve  $C \subset X(p) \times X(p)$ ,*

$$\text{mult}_{\text{CM}^-}(C) = O(p^{-\delta} \text{vol}(C))$$

*Proof.* Take  $d = p^\delta$ . We would like to perform the same trick for the anti-Heegner CM points, and we again partition the points of  $\text{CM}^-$  into

$$T = \text{CM}^- \cap \bigcup_{m < d} \bar{T}_m \quad \text{and} \quad S = \text{CM}^- - T$$

where  $\bar{\cdot}$  denotes complex conjugation on the first factor. For sufficiently small  $\delta' > 0$ , it will still be the case that balls of radius  $\delta' \log p$  around points of  $S$  are

disjoint, but the multiplicity of a curve  $C$  along  $\overline{T}_m$  doesn't make sense, so we adjust the argument slightly:

$$\begin{aligned}
\text{mult}_{\text{CM}^-}(C) &= \sum_{\xi \in S} \text{mult}_{\xi}(C) + \sum_{\xi \in T} \text{mult}_{\xi}(C) \\
&\ll \sinh(\delta' \log p/2)^{-2} \text{vol}(C \cap \cup_{\xi \in S} B(\xi, \delta' \log p)) + \sum_{m < d} \text{mult}_{\text{CM}^- \cap \overline{T}_m}(C) \\
&\ll p^{-\delta'} \text{vol}(C) + \sum_{m < d} \deg T_m \cdot \text{mult}_{\text{CM}^- \cap \overline{\Delta}}(T_m^* C)
\end{aligned} \tag{3}$$

Balls of a fixed small radius  $\epsilon > 0$  ( $\epsilon = 1/10$  is sufficient) around points in  $\text{CM}^- \cap \overline{\Delta}$  are disjoint, and therefore by Theorem 21 we have

$$\begin{aligned}
\text{mult}_{\text{CM}^- \cap \overline{\Delta}}(T_m^* C) &\ll \sum_{\xi \in \text{CM}^- \cap \overline{\Delta}} \text{vol}(T_m^* C \cap B(\xi, \epsilon)) \\
&\ll \text{vol}(T_m^* C \cap B(\overline{\Delta}, \epsilon)) \\
&\ll \sinh(\log p)^{-1} \text{vol}(T_m^* C)
\end{aligned}$$

where we've used Proposition 26 (and Lemma 4) in the last step. Combining this with equation (3), we have

$$\text{mult}_{\text{CM}^-}(C) \ll (p^{-\delta'} + d^3 p^{-1}) \text{vol}(C)$$

and the result follows.  $\square$

**Proposition 29.** *For all sufficiently small  $\delta > 0$ , all sufficiently large  $p$ , and for any non-Hecke curve  $C \subset X(p) \times X(p)$ ,*

$$p \text{mult}_{\text{SBC}}(C) = O(p^{-\delta} \text{vol}(C))$$

*Proof.* Take  $d = p^\delta$ , and again partition the points of SBC into

$$T = \text{SBC} \cap \cup_{m < d} T_m \quad \text{and} \quad S = \text{SBC} - T$$

By Proposition 17, for sufficiently small  $\delta' > 0$  any point in  $X(p) \times X(p)$  is in at most two of the balls  $B(\xi, (1 + \delta') \log p)$  for  $\xi \in S$ . By Theorem 21, for each  $\xi \in \text{SBC}$ ,

$$p^{1+\delta'} \text{mult}_{\xi}(C) = O(\text{vol}(C \cap B(\xi, (1 + \delta') \log p)))$$

and it therefore follows that

$$\begin{aligned}
\text{mult}_{\text{SBC}}(C) &= \sum_{\xi \in S} \text{mult}_{\xi}(C) + \sum_{\xi \in T} \text{mult}_{\xi}(C) \\
&\ll p^{-1-\delta'} \text{vol}(C \cap \cup_{\xi \in S} B(\xi, (1 + \delta') \log p)) + \sum_{\xi \in T} \text{mult}_{\xi}(C) \\
&\ll p^{-1-\delta'} \text{vol}(C) + \sum_{\xi \in T} \text{mult}_{\xi}(C)
\end{aligned} \tag{4}$$

Now for any  $m < d$ , and sufficiently small  $\delta'' > 0$  (independent of  $\delta$ ),

$$\begin{aligned} \text{mult}_{\text{SBC} \cap T_m}(C) &= \sum_{\xi \in \text{SBC} \cap T_m} \text{mult}_{\xi}(C) \\ &\ll \deg T_m \sum_{\xi \in \Delta \cap \text{SBC}} \text{mult}_{\xi}(T_m^* C) \\ &\ll p^{-1-\delta''} \deg T_m \sum_{\xi \in \Delta \cap \text{SBC}} \text{vol}(T_m^* C \cap B(\xi, (1 + \delta'') \log p)) \end{aligned}$$

By 13 part (b), any point within  $(1 + \delta'') \log p$  of two distinct singular bicusps on  $\Delta$  must be within  $\delta'' \log p + O(1)$  of  $\Delta$ , so

$$\begin{aligned} \text{mult}_{\text{SBC} \cap T_m}(C) &\ll p^{-1-\delta''} d^2 \text{vol}(C) + d^2 p^{-1-\delta''} \text{vol}(T_m^* C \cap B(\Delta, \delta'' \log p + O(1))) \\ &\ll p^{-1-\delta''} d^2 \text{vol}(C) (1 + d^2 \cdot O(p^{-1+\delta''/2})) \\ &\ll p^{-1-\delta''+2\delta'} \text{vol}(C) \end{aligned}$$

where the last inequality follows from Proposition 24. Thus,

$$\text{mult}_{\text{SBC}}(C) \ll p^{-1-\delta''+2\delta'+\delta} \text{vol}(C)$$

Choosing  $\delta'' > 2\delta' + \delta$  and combining with equation (4), the result follows.  $\square$

## 6.2. Multiplicity in $(X(p) \times X(p))^2$ .

**Proposition 30.** *For sufficiently small  $\delta > 0$  and sufficiently large  $P$ , any curve  $C \subset (X(p) \times X(p))^2$  no component of which is contained in any diagonal  $\Delta_g$ ,*

$$\sum_g \text{mult}_{\Delta_g} C = O(p^{-\delta} \text{vol}(C)).$$

*Proof.* Let  $\deg_i$  be the degree of the image of  $C$  under the projection  $\pi_i(X(p) \times X(p))^2 \rightarrow X(p) \times X(p)$ , and assume  $\deg_1 \geq \deg_2$ . By Lemma 18, the neighborhoods  $B(\Delta_g, \delta \log p)$  only overlap more than  $O(\log p)$  times within  $O(\delta \log p)$  of a small Hecke curve  $\tau_{g,k}$  with  $k = p^{O(\delta)}$  or within  $(1 + O(\delta)) \log p$  of a point which projects to a singular bicusps along both projections  $\pi_i$ . Let  $E$  be the sum of the volumes of these latter intersections, and let  $T_k^c$  denote the points of  $T_k$  not within  $\delta \log p$  of a singular bicusps. Likewise, denote by  $\tau_{g,k}^c$  the points of  $\tau_{g,k}$  which are not within  $\delta \log p$  of a singular bicusps in either projection  $\pi_i$ .

In the following, we shall use both metrics  $h_{Y(p)}$  and  $h_{X(p)}$ . We specify which metric we are taking the volume with using a subscript. Moreover, when we consider balls in the  $h_{Y(p)}$  metric we write  $B'$  instead of  $B$ . We have

$$\begin{aligned} \sum_g \text{mult}_{\Delta_g} C &\ll p^{-\delta/2} \sum_g \text{vol}_{X(p)}(C \cap B(\Delta_g, \delta \log p)) \\ &\ll E + O(\log p) \cdot p^{-\delta/2} \text{vol}_{X(p)}(C) + p^{-\delta/2} \sum_{k=p^{O(\delta)}} \sum_g \text{vol}_{X(p)}(C \cap B(\tau_{g,k}^c, O(\delta \log p))) \\ &\ll E + o(\text{vol}_{X(p)}(C)) + \deg_1 p^{-\delta/2} \sum_{k=p^{O(\delta)}} \text{vol}_{X(p)}(\pi_1(C) \cap B(T_k^c, O(\delta \log p))) \end{aligned}$$

where the first inequality follows Lemma 22, and the second inequality from Proposition 18. The key observation is that  $T_k^c$  is an étale correspondence with respect to the metric  $h_{Y(p)}$ , and is therefore volume-preserving. Thus, since  $h_{X(p)} \leq h_{Y(p)}$  by Proposition 7, the above is

$$\begin{aligned} &\ll E + o(\text{vol}_{X(p)}(C)) + \deg_1 p^{-\delta/2} \sum_{k=p^{O(\delta)}} \text{vol}_{Y(p)}(\pi_1(C) \cap B'(T_k^c, O(\delta \log p))) \\ &\ll E + o(\text{vol}_{X(p)}(C)) + \deg_1 p^{-\delta/2} \sum_{k=p^{O(\delta)}} \text{vol}_{Y(p)}(T_k^* \pi_1(C) \cap B'(\Delta, O(\delta \log p))) \end{aligned}$$

Now, by Proposition 7, Lemma 24, and Corollary 8 the last term above is bounded by

$$\deg_1 p^{-\delta/2} \sum_{k=p^{O(\delta)}} \left( \text{vol}_{X(p)}(T_k^* \pi_1(C) \cap B(\Delta, O(\delta \log p))) + \frac{1}{p} \text{vol}_{X(p)}(C) \right) \ll p^{O(\delta)-1} \text{vol}_{X(p)}(C).$$

It remains to bound  $E$ . Note that by Proposition 18

$$E \ll \sum_{\xi \in \text{SBC}} \sum_{i=1}^{\log p + O(\delta \log p)} p e^{-i} \text{vol}_{X(p)}(C \cap B(\xi, i)).$$

At the cost of increasing  $\delta$  by a constant factor, by Proposition 24

$$E \ll p^{-\delta} \log p \sum_{\xi \in \text{SBC}} \text{vol}_{X(p)}(C \cap B(\xi, (1+\delta) \log p)).$$

By Proposition 17 the balls  $B(\xi, (1+\delta) \log p)$  are all disjoint for distinct bicusps  $\xi, \xi' \in \text{SBC}$  except when  $\xi, \xi'$  both lie on some  $T_k$  with  $k = p^{O(\delta)}$ . Thus by Proposition 17, Lemma 7, and Corollary 10 we have that

$$\begin{aligned} E &\ll o(\text{vol}_{X(p)}(C)) + p^{-\delta} \log p \sum_{k=p^{O(\delta)}} \sum_{\xi \in \text{SBC} \cap T_k} \text{vol}_{X(p)}(C \cap B(\xi, (1+\delta) \log p)) \\ &\ll o(\text{vol}_{X(p)}(C)) + p^{-\delta} \log p \sum_{k=p^{O(\delta)}} \sum_{\xi \in \text{SBC} \cap T_k} \text{vol}_{Y(p)}(C \cap B(\xi, (1+\delta) \log p)) \\ &\ll o(\text{vol}_{X(p)}(C)) + p^{-\delta} \log p \sum_{k=p^{O(\delta)}} \sum_{\xi \in \text{SBC} \cap \Delta} \text{vol}_{Y(p)}(T_k^* C \cap B(\xi, (1+O(\delta)) \log p + O(1))) \\ &\ll o(\text{vol}_{X(p)}(C)) + p^{-\delta} \log p \sum_{k=p^{O(\delta)}} \sum_{\xi \in \text{SBC} \cap \Delta} \text{vol}_{Y(p)}(T_k^* C \cap B(\Delta, O(\delta) \log p + O(1))) \\ &\ll o(\text{vol}_{X(p)}(C)) + p^{-\delta} \log p \sum_{k=p^{O(\delta)}} \sum_{\xi \in \text{SBC} \cap \Delta} \text{vol}_{X(p)}(T_k^* C \cap B(\Delta, O(\delta) \log p + O(1))) \\ &\ll o(\text{vol}_{X(p)}(C)) + p^{O(\delta)-1} \text{vol}_{X(p)}(C). \end{aligned}$$

Taking  $\delta$  sufficiently small establishes the proposition.  $\square$

## 7. PROOF OF THE MAIN THEOREM

### 7.1. Gonality.

Recall that for a proper algebraic curve  $C$ , the gonality  $\text{gon}(C)$  of  $C$  is the smallest integer  $d$  for which there is a degree  $d$  map  $C \rightarrow \mathbb{P}^1$ . For example,  $\text{gon}(C) = 1$  if and only if  $C \cong \mathbb{P}^1$ , and  $C$  is said to be hyperelliptic if  $\text{gon}(C) = 2$ . In particular, every genus 1 (or 2) curve is hyperelliptic, but it is easy to show that there are hyperelliptic curves of every genus  $g > 0$ .

In general the gonality of a curve  $C$  is difficult to compute, but it is always bounded in terms of the genus  $g = g(C)$ ,

$$\text{gon}(C) \leq g(C) + 1$$

by Riemann–Roch. In fact it is well known from Brill–Noether theory that

$$\text{gon}(C) \leq \left\lfloor \frac{g(C) + 3}{2} \right\rfloor \quad (5)$$

is a strict inequality in the sense that a generic curve  $C$  will achieve the bound in (5). Theorems 1 and 2 follow immediately from the following

**Theorem 31.** *For any  $B > 0$ , there exists  $C_B > 0$  such that for any smooth curve  $V$  of gonality  $n < B$ , any nonconstant map  $V \rightarrow Z(p)$  factors through a Hecke curve, provided  $p > N$ .*

*Proof.* Suppose not, so that for arbitrarily large  $p$  we have a smooth curve  $V \rightarrow Z(p)$  of bounded gonality  $n$  not factoring through a Hecke curve, which we may assume is degree 1 onto its image. The degree  $n$  linear system on  $V$  gives a map  $\varphi : \mathbb{P}^1 \rightarrow \text{Sym}^n Z(p)$  which is also degree 1 onto its image. Let  $\psi : C \rightarrow (X(p) \times X(p))^n$  be the normalization of an irreducible component of the pullback of  $\varphi$  to  $(X(p) \times X(p))^n$ , and  $\alpha : C \rightarrow \mathbb{P}^1$  the resulting map:

$$\begin{array}{ccc} C & \xrightarrow{\psi} & (X(p) \times X(p))^n \\ \alpha \downarrow & & \downarrow \\ \mathbb{P}^1 & \xrightarrow{\varphi} & \text{Sym}^n Z(p) \end{array}$$

The Galois group  $G$  of  $(X(p) \times X(p))^n$  over  $\text{Sym}^n Z(p)$  is an extension

$$1 \rightarrow G(p)^n \rightarrow G \rightarrow S_n \rightarrow 1$$

Let  $H \subset G$  be the stabilizer of  $C$ ,  $G_i \subset G$  the subgroup of elements that act trivially on the  $i$ th factor of  $(X(p) \times X(p))^n$ , and  $H_i = H \cap G_i$ . Thus, the normalization of the image of  $C$  under the projection  $\pi_i$  is identified with  $C/H_i$ .

We will bound the degree  $\text{Ram}_\alpha$  of the ramification divisor of  $\alpha$ . A point  $Q \in C$  ramifies only if  $\xi = \psi(Q)$  is in

$$\Delta_g^{ij} := \{(x_1, y_1, \dots, x_n, y_n) \mid x_i = gx_j, y_i = gy_j\} \subset (X(p) \times X(p))^n$$

for some  $i, j \in \{1, \dots, n\}$  and  $g \in G(p)$ , or if there exists some  $i \in \{1, \dots, n\}$  such that  $\pi_i(\xi)$  is either a singular bicuspid, Heegner CM point or an anti-Heegner CM point. The analytic local stabilizer of  $Q$  is a cyclic subgroup of  $H$  and therefore of order  $O(p)$ . The ramification index of  $\alpha$  at  $Q$  is then also  $O(p)$ , and in fact



if  $\xi$  does not project to a singular bicuspid in any projection, the index is  $O(1)$  (bounded by  $6n!$ ). It follows therefore that

$$\text{Ram}_\alpha \ll p \sum_i |(\pi_i \circ \psi)^{-1}(\text{SBC})| + \sum_i |(\pi_i \circ \psi)^{-1}(\text{CM}^+ \cup \text{CM}^-)| + \sum_{i,j,g} |(\pi_{ij} \circ \psi)^{-1}(\Delta_g)|$$

Our ramification estimates then give us control over these three terms:

- (1)  $|\psi^{-1}(\xi)|$  is bounded by the multiplicity  $\text{mult}_\xi \psi(C)$ , so

$$|(\pi_i \circ \psi)^{-1}(\text{SBC})| \leq |H_i| \text{mult}_{\text{SBC}} \pi_i \circ \psi(C)$$

But  $\pi_i \circ \psi(C)$  is normalized by  $C/H_i$ , so  $\text{vol}(C) = |H_i| \text{vol}(\pi_i \circ \psi(C))$ , and by Proposition 29 we have

$$p |(\pi_i \circ \psi)^{-1}(\text{SBC})| = o(\text{vol}(C)).$$

- (2) Likewise,

$$|(\pi_i \circ \psi)^{-1}(\text{CM}^+ \cup \text{CM}^-)| \leq |H_i| \text{mult}_{\text{CM}^\pm} \pi_i \circ \psi(C) = o(\text{vol}(C))$$

by Propositions 27 and 28.

- (3) Finally, the projection of  $C$  to the  $ij$ th factor has degree  $|H_i \cap H_j|$  over its image, so we similarly have

$$\sum_g |(\pi_{ij} \circ \psi)^{-1}(\Delta_g)| \leq |H_i \cap H_j| \sum_g \text{mult}_{\Delta_g} \pi_{ij} \circ \psi(C) = o(\text{vol}(C))$$

by Proposition 30.

Thus,  $\text{Ram}_\alpha = o(\text{vol}(C))$ , and Riemann–Hurwitz applied to  $\alpha$  yields

$$g(C) = o(\text{vol}(C))$$

However, if  $d$  is the largest degree of the projections  $C \rightarrow X(p)$ , so that

$$\text{vol}(C) \leq 2nd \text{vol}(X(p))$$

then Riemann–Hurwitz applied to this projection yields

$$g(C) \gg dg(X(p))$$

and the above three equations would give  $g(X(p)) = o(\text{vol}(X(p)))$  which is false, since  $g(X(p))$  and  $\text{vol}(X(p))$  are asymptotically proportional to  $p^3$ , as  $\text{vol}(X(p)) = |\text{PSL}_2(\mathbb{F}_p)| \text{vol}(X(1)_p)$ . This contradiction establishes the theorem.  $\square$

## 7.2. Applications.

By the remarks preceding Theorem 31, we obtain as a corollary the following weaker result:

**Corollary 32.** *For any  $B > 0$ , there exists  $N > 0$  such that for any smooth curve  $V$  of genus  $g < B$ , any nonconstant map  $V \rightarrow Z(p)$  factors through a Hecke curve, provided  $p > N$ . In particular, for  $p$  sufficiently large, every rational or elliptic curve on  $Z(p)$  is a Hecke curve.*

In particular,  $Z(p)$  has no rational or elliptic curves other than Hecke curves, thus answering a question first posed by Kani and Schanz [KS98].

The surface  $Z(p)$  has cyclic quotient singularities, each locally analytically isomorphic to the quotient of  $\mathbb{C}^2$  by  $\mathbb{Z}/n\mathbb{Z}$  acting by  $i \cdot (z, w) = (\zeta^i z, \zeta^{ai} w)$  for a primitive  $n$ th root of unity  $\zeta$  and some  $0 < a < p$ . For CM points  $n = 2$  or  $3$ , while for singular bicusps  $n = p$ . The minimal resolution  $q : \tilde{Z}(p) \rightarrow Z(p)$  resolves such a singular point into a chain of smooth rational curves whose intersection form is determined by the continued fraction expansion of  $\frac{n}{a}$ .

Corollary 32 also resolves a conjecture of Hermann [Her91]:

**Corollary 33.** *For all sufficiently large  $p$ , the minimal model of  $Z(p)$  is obtained from  $\tilde{Z}(p)$  by blowing down “known” curves, i.e. by blowing down strict transforms of Hecke curves and curves contracted by  $\tilde{Z}(p) \rightarrow Z(p)$ .*

*Proof.* By Corollary 32 all rational curves in  $\tilde{Z}(p)$  are of this type. □

## REFERENCES

- [Abr96] Dan Abramovich. A linear lower bound on the gonality of modular curves. *Internat. Math. Res. Notices*, (20):1005–1011, 1996.
- [BS94] P. Buser and P. Sarnak. On the period matrix of a Riemann surface of large genus. *Invent. Math.*, 117(1):27–56, 1994. With an appendix by J. H. Conway and N. J. A. Sloane.
- [BT13] B. Bakker and J. Tsimerman. On the Frey-Mazur conjecture over low genus curves. [arXiv:1309.6568](https://arxiv.org/abs/1309.6568), 2013.
- [Car01] David Carlton. Moduli for pairs of elliptic curves with isomorphic  $N$ -torsion. *Manuscripta Math.*, 105(2):201–234, 2001.
- [Fis11] T.A. Fisher. On families of  $n$ -congruent elliptic curves. [arXiv:1105.1706](https://arxiv.org/abs/1105.1706), 2011.
- [Her91] C. F. Hermann. Modulflächen quadratischer Diskriminante. *Manuscripta Math.*, 72(1):95–110, 1991.
- [HT02] J. Hwang and W. To. Volumes of complex analytic subvarieties of Hermitian symmetric spaces. *American Journal of Mathematics*, 124(6):1221–1246, 2002.
- [HT12] J. Hwang and W. To. Injectivity radius and gonality of a compact Riemann surface. *American Journal of Mathematics*, 134(1):259–283, 2012.
- [KS98] E. Kani and W. Schanz. Modular diagonal quotient surfaces. *Mathematische Zeitschrift*, 227(2):337–366, 1998.
- [MG78] B. Mazur and D. Goldfeld. Rational isogenies of prime degree. *Inventiones mathematicae*, 44(2):129–162, 1978.

B. BAKKER: COURANT INSTITUTE OF MATHEMATICAL SCIENCES, NEW YORK UNIVERSITY, 251 MERCER ST., NEW YORK, NY 10012

*E-mail address:* bakker@cims.nyu.edu

J. TSIMERMAN: MATHEMATICS DEPARTMENT, HARVARD UNIVERSITY, 1 OXFORD STREET, CAMBRIDGE, MA, 02138

*E-mail address:* jacobt@math.harvard.edu